Distribution Agreement

Signature:

______________________________ April 17, 2017

Dante G. Bugli Date

Multiple Imputation Method in SAS Exemplified through a Case Study of Programmatic Data from Emergency Nutrition Programs

By

Dante G. Bugli

Master of Public Health

Global Epidemiology

Carlos Navarro-Colorado, MD PhD

Committee Member

Kristin Wall, PhD

Committee Chair

Multiple Imputation Method in SAS Exemplified through a Case Study of Programmatic Data from Emergency Nutrition Programs

By

Dante G. Bugli

Bachelor of Science in Brain Behavior and Cognitive Science

University of Michigan

2011

Thesis Committee Chair: Kristin Wall, PhD

Committee Member: Carlos Navarro-Colorado, MD PhD

An abstract of

a thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Public Health

in Global Epidemiology

2017

Abstract

Multiple Imputation Method in SAS Exemplified through a Case Study of Programmatic Data from Emergency Nutrition Programs

By Dante Bugli

Background. Missing data is a problem that all researchers encounter. Historically applied imputation methods expose a study to bias while advanced statistical methodology called multiple imputation (MI) method introduces the smallest amount of bias. Drawing upon a complex theoretical basis, statistical software responded accordingly by providing a sound and rapid application of MI. Few resources exist detailing the application of the method.

Objective. This paper provides a brief explanation of the foundations of MI method and applies it as a sensitivity analysis of a study implemented across three countries. By comparing results of model selection from both analyses, factors of significant impact on programmatic success can be more clearly identified.

Methods. Using a dataset of information from an exit questionnaire of a supplemental feeding program (SFP) implemented in emergency settings, MI was applied to artificially complete the dataset. Bivariate and multivariate regression were used to determine appropriate models to identify important factors that would lead to a patient defaulting from the program.

Results. Missing data was a large problem in this case study’s dataset with variables ranging from 14% to 52% missing. MI completed the datasets and produced 10 imputed datasets for multivariate analysis. Models selected based on the imputed datasets were not entirely identical to those from the original analysis but reflected similar adjusted odds ratios with higher precision for those that coincided.

Conclusions. MI was valuable as a sensitivity analysis to identify important modifiable factors to decrease program defaulting. By identifying factors that were significantly influencing or impeding participants’ abilities/desire to remain in the SFP future programming may be improved. This paper shows that applying MI to categorical datasets can still confirm the results of a primary analysis and aid in targeting key factors.

Keywords: Multiple Imputation; Missing Data; Statistical Analysis Systems (SAS); Supplemental Feeding Programs (SFP);

Multiple Imputation Method in SAS Exemplified through a Case Study of Programmatic Data from Emergency Nutrition Programs

By

Dante G. Bugli

Bachelor of Science in Brain Behavior and Cognitive Science

University of Michigan

2011

Thesis Committee Chair: Kristin Wall, PhD

Committee Member: Carlos Navarro-Colorado, MD PhD

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Public Health

in Global Epidemiology

2017

Table of Contents

INTRODUCTION

Missing data is a problem across all forms of research. In cross-sectional to randomized control trials, despite a strong emphasis on data collection methods, it is nearly impossible to avoid missing data. Depending on one's study design, missing data can be caused by a variety of factors. Defaulting or censoring in longitudinal studies result in missing data. In other studies, missing data on the severity of a patient's condition can be caused by the severity itself preventing the patient from reporting or visiting the clinic. The reasons behind missing data become more important when choosing the appropriate method for adjusting for the missingness or the general analysis because each strategy is based upon assumptions of the mechanism causing the missing data (Pigott, 2001).

When considering the mechanism for missing data, those who collected the data could often explain the reasoning behind missing data, but rarely is the data collector and the data analyst the same person. From the analyst's perspective, the reason an individual item is missing is not as important as the systematic lack of data. In the case that there is missing data throughout the dataset in no consistent or systematic pattern, one can make the argument that the missingness mechanism is *Missing Completely at Random* (MCAR) (Rubin, 1996). This being the ideal scenario though rarely the case, one must address problems when the data is *Missing at Random* (MAR) or has *Missingness that depends on unobserved predictors* (Little and Rubin, 2014). Delineating the difference between these two cases is empirically impossible; therefore allowing analysts to group the two together as data missing dependent on a variety of other variables but not the missing variable itself. Finally, the most difficult case is when data is *Missing Not at Random* (MNAR) referring to when the probability of an item being missed is dependent on the item itself (Berglund and Heeringa, 2014). Analysts often look at a dataset and describe the pattern of missingness to choose the appropriate methods for addressing the missing data.

As statistical methods have advanced, data analysts are provided a breadth of tools to address the problems that missing data can cause during analysis. The capacity of conventional methods met the needs of analysts at a time when the understanding and approach for dealing with missing data was less advanced. Methods such as complete-case analysis or available-case analysis leave out large chunks of data which may introduce bias (Little and Rubin, 2014). For example, in using the complete-case approach, where one discards any subject with a missing item, an analyst runs the risk of introducing a bias against the group with missing data. The group that gets discarded may have an underlying relationship to the outcome or to each other that is the cause of the missing items meaning the analyst is disregarding a pattern of interest. Statistically informed methods such as *mean imputation* or *last observation carried forward (LOCF)*, attempt to impute the missing item informed by the surrounding responses within the dataset (de Goeij, 2013). These, too, may cause bias. The *mean imputation* approach will create a middling-effect by implying that the distribution is strongly centered on the mean. The insufficient management of missing data meant that further methods were needed (Soley-Bari, 2013).

The methods noted above are a small number of the most commonly used methods among a long list of tactics to address missing data. Complete-case analysis is the default approach for multiple procedures within statistical software programs such as SAS and R (de Goeij, 2013). Statistician Donald Rubin was one of the first to push for the evolution and increased usage of multiple imputation as a means for handling missing data (Rubin, 1996). Frequent implementation of the method lagged behind the development of the statistical validity of the method due to its complexity and unavailability in common programming software. It was only in 2004 that the SAS Version 9.0 included statements that allowed for quick analysis using PROC MI and PROC MIANALYZE (Yuan, 2010).

This paper will use multiple imputation of categorical data in a study of nutritional program data. The data used for this case study is pulled from an exit survey used to assess the reasons participants default out of supplemental feeding programs in emergency settings. While working for the Emergency Nutrition Network (ENN), the head researcher, Dr. Carlos Navarro-Colorado, sought to establish what common factors would lead to a participant to leave the program as a means to improve the delivery of life-saving nutritional programs in at-need communities (2007).

Implemented across three countries (Chad, Sudan, and Kenya), missing data became apparent upon initial analysis of the entrance survey (Schroeder, *unpublished*). Multiple imputation performed on the data from the exit survey will allow for comparison between the model selection completed using the imputed dataset and the model using complete-case analysis (Palmer, *unpublished*). This case study will investigate whether the missing data biased the results.

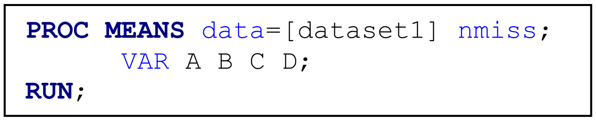
METHODS

Preparation for Imputation

When preparing for multiple imputation (MI), there are several questions prior to analysis to answer: (1) what variables need to be included, (2) what is the extent and pattern of missingness, and finally, (3) what type of variables are included i.e. nominal, continuous, etc. (Berglund and Heeringa, 2014)?

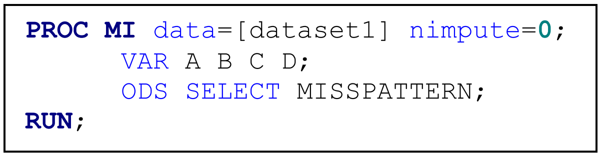
When assessing the dataset prior to imputation, the analyst must make intentional decisions as to which variables will be included in the imputation. The first variable(s) chosen must be the primary outcome(s) of interest. Following this, any variables the analyst is wishing to impute that are of importance whether due to their effects on the outcome or interest to the researcher are considered. Finally, additional auxiliary variables of the "to-be imputed" variables may be included (Soley-Bari, 2013). Auxiliary variables can be chosen based on a known relationship or whether they predict the missingness of the variables that will be imputed (Berglund and Heeringa, 2014). Arguments over the number of auxiliary variables have gone back and forth as to the validity they provide and at what cost, statistically speaking. Further investigation has shown that the benefit of including more auxiliary variables outweighs their statistical cost (DiazOrdaz, 2016).

Quantifying the amount of missing data and the pattern in which the items are missing must be identified and considered before imputing. Statistical software allows for several strategies to report the number of missing items by variable. A typical procedure such as PROC MEANS can be used in SAS with an added *nmiss* option in the procedure statement to provide a breakdown of how many items are missing per variable.



```
PROC MEANS data=[dataset1] nmiss;
       VAR A B C D;
RUN;
```

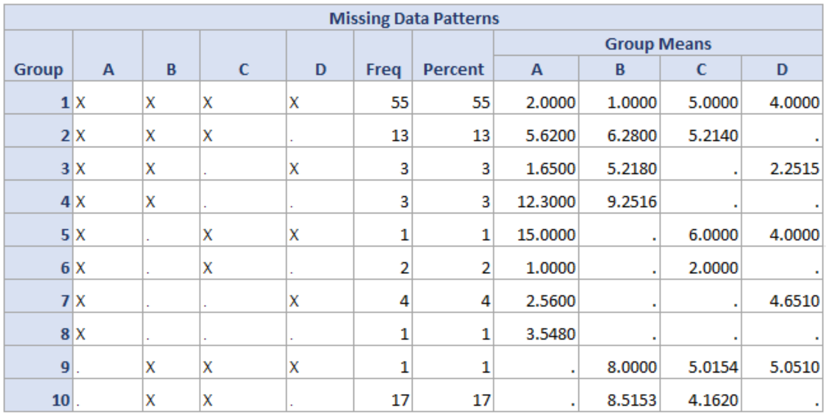
Figure 1: Example of SAS (Version 9.4) code used to produce the frequency of missing items by variable.



```
PROC MI data=[dataset1] nimpute=0;
       VAR A B C D;
       ODS SELECT MISSPATTERN;
RUN;
```

Figure 2: Example of SAS (Version 9.4) code used to produce the missing pattern figure as seen in Table 1.

The PROC MI statement can be used in this step to get a more global view of the missingness patterns. By running the statement without any imputations, as indicated by nimpute=0, and including the ODS SELECT MISSPATTERN statement, SAS will print out the different combinations of missing and non-missing data including the number of times each pattern arises and the percentage of subjects that pattern represents (Berglund and Heeringa, 2014). The example below shows that only 55% of the total sample size has complete data whereas 17% are missing variable A and variable D.

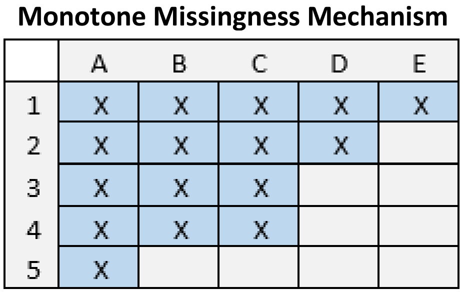


Missing Data Patterns

							Group Means			
Group	A	B	C	D	Freq	Percent	A	B	C	D
1	X	X	X	X	55	55	2.0000	1.0000	5.0000	4.0000
2	X	X	X	.	13	13	5.6200	6.2800	5.2140	.
3	X	X	.	X	3	3	1.6500	5.2180	.	2.2515
4	X	X	.	.	3	3	12.3000	9.2516	.	.
5	X	.	X	X	1	1	15.0000	.	6.0000	4.0000
6	X	.	X	.	2	2	1.0000	.	2.0000	.
7	X	.	.	X	4	4	2.5600	.	.	4.6510
8	X	.	.	.	1	1	3.5480	.	.	.
9	.	X	X	X	1	1	.	8.0000	5.0154	5.0510
10	.	X	X	.	17	17	.	8.5153	4.1620	.

Figure 3: Adapted output from SAS (Version 9.4) of the Missing Data Patterns produced by the PROC MI statement showing 10 unique patterns among four variables with their associated frequencies and group means.

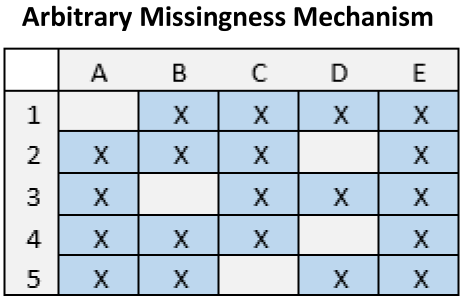
Using PROC MI and the Missing Data Pattern output is important in determining the pattern of missingness or the missingness mechanism. Identifying the missingness mechanism will allow us to choose an appropriate imputation modeling approach because the inferences made in imputation depend on the process that leads to the missing data (DiazOrdaz, 2016). There are two categories of missingness mechanisms: monotone and arbitrary. A monotone missingness mechanism is characterized by subjects who have missing data from one point through the rest of dataset. This is most commonly found in longitudinal studies where a participant may have been lost to follow-up before the end of the study. Frequently, datasets will have a more generalized, non-systematic missingness pattern which is called arbitrary (Berglund and Heeringa, 2014). Though evident when using the PROC MI statement, this determination will greatly influence the core imputation steps. Accounting for monotonic missing data patterns requires a much less robust analysis, whereas several options have been developed for handling arbitrary missing patterns. These approaches will be explored in the next section.



Monotone Missingness Mechanism

	A	B	C	D	E
1	X	X	X	X	X
2	X	X	X	X	
3	X	X	X		
4	X	X	X		
5	X				

Figure 4: Illustration of typical monotone missingness mechanism.



Arbitrary Missingness Mechanism

	A	B	C	D	E
1		X	X	X	X
2	X	X	X		X
3	X		X	X	X
4	X	X	X		X
5	X	X		X	X

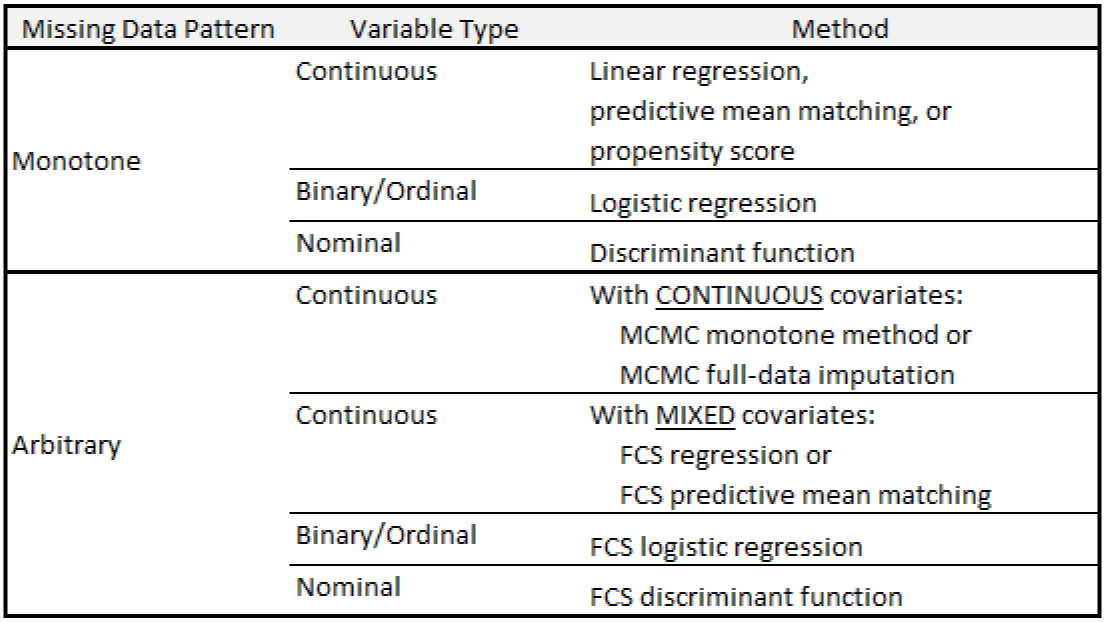
Figure 5: Illustration of typical arbitrary missingness mechanism.

The final step in preparing for imputation is to identify the types of variables that will be imputed. Within the MI technique, there are several specific approaches that have been developed and determining which is the most appropriate for a given variable depends on the type of variable that is being imputed (Rubin, 1996). Similar to differing approaches to modeling, a continuous variable cannot be imputed using the same technique as an ordinal variable.

In addition to these preparatory steps, there are two underlying properties of the data that should be checked. Continuous variables that are notably skewed should be considered for transformation. A strong skew can misconstrue the parameter estimates during multiple imputation. This is particularly important in using fully conditional specific (FCS) functions which will be discussed below (Messer and Natarajan, 2008). Secondly, we return to the concept of data being Missing at Random (MAR) or Missing Completely at Random (MCAR) versus Missing Not at Random (MNAR). By stating that the missing data is MCAR or MAR, the following analysis steps will be simpler because the analyst is justified in ignoring the missingness mechanism (Soley-Bori, 2013). The distinction between all three is difficult to establish and nearly impossible to empirically define (Pigott, 2001). In most cases, datasets are not solely one of those definitions. It is more probable that a dataset is a mix of MNAR and MAR, therefore analysts can continue under the assumption that the majority is MAR allowing for the methods used in MI (de Goeij, 2013). MCAR is seen as a special case within the larger category of MAR which is rarely achieved. More detail into how to treat MNAR datasets can be found in the full text of Statistical Analysis with Missing Data by Little and Rubin (2014).

Imputation Phase

Imputation may be one option of many when analyzing missing data, but as technical programming has advanced giving researchers more access to the method, its use has increased for several reasons. The assumptions of many conventional methods limit their usage and may introduce bias (de Goeij, 2013). Additionally, the MI method is more flexible to a variety of variables and can handle multivariate analysis. When employed as a method in a validation study, the MI performed competitively when compared to other popular methods of estimation such as Maximum-Likelihood (ML) with Expectation-Maximization (EM) (Messer and Natarajan, 2008).



Missing Data Pattern	Variable Type	Method
Monotone	Continuous	Linear regression, predictive mean matching, or propensity score
	Binary/Ordinal	Logistic regression
	Nominal	Discriminant function
Arbitrary	Continuous	With CONTINUOUS covariates: MCMC monotone method or MCMC full-data imputation
	Continuous	With MIXED covariates: FCS regression or FCS predictive mean matching
	Binary/Ordinal	FCS logistic regression
	Nominal	FCS discriminant function

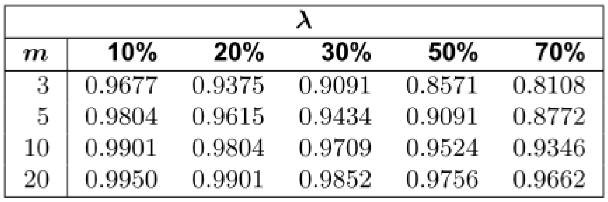
Figure 6: Imputation modeling method selection as determined by missing data pattern and variable type. Adapted from figure found in Berglund and Heeringa (2014).

The initial phase of MI is the imputation or *i-phase* wherein the imputation models are defined and the separate iterations of datasets filled with plausible values are created. It is within this phase, and the model definition in particular, that the method's flexibility is applied. By accounting for the information noted in the preparation for imputation, each variable that is chosen for imputation will require a specific model used under particular conditions. As discussed above, monotone and arbitrary datasets will be treated differently. Those that are monotone allow for more straightforward modeling strategies such as linear regression, predictive mean matching, or logistic regression (Figure 6). The strategy is chosen based on the type of variable that is being imputed (Berglund and Heeringa, 2014).

The Markov chain Monte Carlo (MCMC) method first allowed for treatment of continuous variables that were arbitrarily missing to be analyzed by assuming a multivariate normality distribution for the missing variables (Schafer, 1997). This assumption produces the plausible values to fill the missing data through an "algorithm that alternates between estimating the parameters of the multivariate normal distribution and producing imputed values from the appropriate posterior predictive distributions" (Romaniuk, Patton, and Carlin, 2014). By setting this posterior predictive distribution, the plausible values are sampled multiple through "burn-in repetitions" that continue to a set number or until convergence. The "burn-in repetitions" can affect the results of the imputation and as such may be manipulated by adding the NBITER option into the PROC MI statement. The multivariate normal assumption applies to limited scenarios of variables that are continuous and being modeled based on other continuous variables.

As noted in Figure 6 above (reimaged from Berglund and Heeringa, 2014), most scenarios will require the use of fully conditional specific (FCS) methods. To address the more likely scenario of dealing with continuous and categorical variables within the same model, Buuren, Boshuizen, and Knook introduced the idea of multiple imputation by chained equations (1999). By this process, an iterative sequence of draws will simulate draws from a joint posterior distribution of parameters. The iterations will continue until convergence. The sequence is not clearly based on the Bayesian inference frameworks, therefore the distributions mentioned above are not established but simulated (Yuan, 2010).

Though most typical statistical software programs will have a method for multiple imputation, for the case study presented in this paper, the author researched and used SAS V9.4 methods exclusively. As previously mentioned, it is only within the past 10 years that this procedure has become standardized by a formal procedure in SAS (Yuan, 2010). Given such novelty to the technique, it is pertinent to include a small appendix of the approach used (APPENDIX I). There are many other options within the procedure that will not be mentioned in this paper, but this may serve as a reference for basic level multiple imputation. When coding the analysis, the indication of the number of imputations is made. Choosing a defensible number of imputations continues to be debated among statisticians, but the typical number falls between 5 and 30 (Romaniuk, Patton, and Carlin, 2014; Berglund and Heeringa, 2014). This choice should be based on the percentage of missingness in the dataset. Yuan, of the SAS Institute, presented a table comparing the number of imputations, m, compared the percent of missing values and their resulting relative efficiency, lambda (2010). The second key piece of this coding is the imputation model itself. The methods may be unique to each variable to be imputed. There is no requirement that the imputation models resemble the analysis model, but it should inform the decisions of which variables to include.



	λ				
m	10%	20%	30%	50%	70%
3	0.9677	0.9375	0.9091	0.8571	0.8108
5	0.9804	0.9615	0.9434	0.9091	0.8772
10	0.9901	0.9804	0.9709	0.9524	0.9346
20	0.9950	0.9901	0.9852	0.9756	0.9662

Figure 7: Multiple imputation efficiency by percentage missing as calculated by the formula proposed by Rubin (1987) and displayed in this table by Yuan (2010).

Analysis and Pooling Phase

Multiple imputation performed in SAS is done in three phases: the imputation phase, the analysis phase, and the pooling phase. The product of the imputation phase should be the same number of complete datasets as the number of imputations that were chosen in the PROC MI statement. Each one of these datasets no longer has any missing data points which were all filled in with plausible values based on the distribution of available and missing data (Berglund and Heeringa, 2014). The subsequent two phases are done nearly simultaneously.

When analyzing the datasets, we recall typical regression analysis to produce parameter estimates. These models are not beholden to the original model of analysis (prior to imputation) and may include any set of the variables used in the imputation process (DiazOrdaz, 2016). Any regression technique may be employed for this step; common choices would be PROC LOGISTIC, PROC SURVEYLOGISTIC, or PROC REG. In running a PROC MI statement, SAS is implicitly creating a new variable for your now larger dataset named "_imputation_". A BY statement is required in the analysis step to indicate that each imputed dataset is treated separately (Berglund and Heeringa, 2014). The important output from this step is the set of parameter estimates and their associated standard errors which can be sent to a new dataset. Each survey procedure will have a slightly different syntax for the needed information which can be clarified further in most SAS guides. An example of this syntax can be found in Appendix II.

Multiple imputation is completed by using the PROC MIANALYZE to combine the multiple sets of parameter estimates. The procedure takes into account the parameters estimates as well as their associated standard errors. It is through this final step that we obtain a model based on all of the imputed datasets. Being a fairly novel procedure in SAS, a third appendix demonstrates an example of what options are available and which are required within the PROC MIANALYZE statement (APPENDIX III).

ANALYSIS

Ethics statement

The study being used to illustrate the application of MI was approved by the Director of the CNNTA (Nutrition Department for the Chad Ministry of Health), Nutrition Manager in the Kenyan Division of Nutrition, and the General Secretary in the Sudan Ministry of Public Health, and all participants provided oral informed consent.

Sample

The dataset of interest in this study is from the Defaulting and Access Study run by Dr. Carlos Navarro-Colorado while working for Emergency Nutrition Network (ENN) in collaboration with Jeremy Shoham and Frances Mason. The study focuses on the reasons why a participant may default out of a supplemental feeding program. In the context of this study, a defaulter was defined "a beneficiary that is lost to the programme before reaching discharge criteria, and whose actual status (dead, recovered, other) is not known" (Navarro-Colorado, Mason and Shoham, 2010). As patients entered and exited the study, they were administered a questionnaire.

The full dataset contained 2,003 observations, collected by programs run by Action Against Hunger, Save the Children, and Concern Worldwide, were cut down to 1,792 by removing any that were missing the outcome variable. These observations were distributed between three countries: Chad (687), Kenya (297) and Sudan (808). The dataset being analyzed in this case is that of the exit survey, which evaluated participant's perceptions of the study, experiences when attending the clinic, and logistics related to their attendance at the clinic (Navarro-Colorado, Mason and Shoham, 2010). Missing data was seen for most variables and was artificially inflated by recoding answers that were “Other” or “Unknown” as missing before analysis began. The subset of variables chosen for this analysis mirrored that of the primary analysis done by Palmer (2017). The number of variables was limited to focus on the data that relates to a patient’s ability and desire to remain in the study. A total of 61 variables, all categorical, were included in the initial analyses. Before beginning imputation, correlation tests were run on select variables that were not already in the chosen set. Based on a strong correlation coefficient, auxiliary variables would be included within the imputation models.

The SAS procedures were used to quantify the amount of missing items per variable as well as to produce the missingness patterns. These outputs would illustrate which kind of missingness mechanism is affecting the dataset. These results were then used to inform imputation model decisions. Prior to moving forward with any analysis, preliminary data led to the key assumption that the data missing was missing at random (MAR). The author acknowledges that while some trends in the missingness may appear missing not at random (MNAR), the large majority of the missingness was MAR allowing for multiple imputation method to be applied.

Imputation Phase

When designing the imputation models for the imputation phase, all variables that were missing items were considered for imputation, despite large percentages of missing data for some variables (up to 52%). For each variable’s imputation model, the full set of variables were used, including the outcome, gender and country data. A specific seed (1001) which is used as the starting point for random number generation was chosen to ensure repeatability. Ten imputations were chosen to reflect the high amount of missingness but the large amount of available auxiliary variables led to a lower amount of imputations than some literature would suggest.

Since all variables being imputed were categorical variables, with suspected arbitrary missingness, only fully conditional specific (FCS) model approaches were used. Literature supports that this is the most appropriate method for such data (Berglund and Heeringa, 2014). The majority of the chosen variables were coded as binary though three were nominally coded categorical variables with multiple levels. For those binary variables, FCS logistic regression was used as an imputation model while those with multiple levels were imputed using the FCS discriminatory function. Examples of both model designs can be seen in the box below.

Analysis and Pooling Phase

Following imputation, we rely on standard survey methods to analyze the imputed datasets. The results of those analyses are then pooled using the final multiple imputation analysis step. The near simultaneous nature of these two steps means one must begin to consider an applicable model selection approach. To ensure comparability between this case study and the previously completed analysis (Palmer, 2017), identical approaches were used in selecting a model and eliminating unnecessary variables from the model. Each variable was analyzed individually for its association with the outcome. If found to be insignificant (p>=0.05) under a bivariate analysis, the variable would not be considered in the multivariate analysis. Using logistic regression, parameter estimates and their associated covariances and/or standard errors were extracted for analysis. Once an initial multivariate regression was completed, insignificant factors were dropped and a primary multivariate model was run to obtain more accurate estimates. Reported in the results are unadjusted and adjusted odds ratios (OR, aOR respectively) and their associated 95% confidence limits. The Statistical Analysis System (SAS) 9.4 English version was used for all analyses except the multivariate analysis which was run on SAS 9.3 English version.

RESULTS

Analysis of Missingness

To begin, responses indicating “other” and “I don’t know” were recoded as missing before the missingness analysis was run. By variable, the largest percentage missing was 52% missing for all of the responses to each sub-question (*n* = 13) of “How could things be improved at the SFP?” Only sub-questions of “Did you experience any of the following during the time the child was following the nutrition programme?" reported 0% missing. All other variables (*n* = 17) analyzed fell between 17% and 21% missing.

An analysis of the missingness patterns showed that there are 70 unique patterns of missingness. The most frequent pattern was a completed dataset (38.06%) followed by the pattern that represents a complete dataset except for all sub-questions to "How could things be improved at the SFP?" (33.65%) which was identified previously as having the highest missing percentage. The final pattern of interest was that missing all data except the complete set of sub-questions from "Did you experience any of the following during the time the child was following the nutrition programme?" which represented 16.8% of the patterns. The rest of the patterns occurred less than 2% of the time wherein they would be missing random items along with commonly missing items. An annotated graphic representation of the missingness patterns (Table 1) verifies that the dataset consists of arbitrarily missing data as opposed to monotone.

Imputation

Imputation models were set for all variables with >0% of missing data run separately by country. Auxiliary variables tested by the Pearson Correlation test proved insignificant for inclusion. Bivariate analysis for all variables being considered were performed and those that were significant (alpha level of 0.05) were included in the multivariate analysis.

The program in Chad produced a model (Table 2) showing a significantly increased odds of defaulting out of the program if the family was "more busy than usual" in the past 3 months (aOR: 3.2; 95% CI: 1.67, 6.15), if they believed the program would be improved by being "[asked] to come less often" (aOR: 3.15; 95% CI: 1.44, 6.87), and if they thought that the staff was giving out the wrong ration (aOR: 27.87; 95% CI: 7.34, 100.38). Conversely, families that were "less busy than expected" in that year compared to other years (aOR: 0.28; 95% CI: 0.11, 0.69), described their experience as good (aOR: 0.032; 95% CI: 0.006, 0.17), and experienced the illness of the participating child (aOR: 0.36; 95% CI: 0.19, 0.69), that they didn't feel the child's health was improving (aOR: 0.24; 95% CI: 0.11, 0.55), that the SFP had lost or withdrawn their card for participation (aOR: 0.03; 95% CI: 0.01, 0.10), or that the child appeared to be recovered (aOR: 0.12; 95% CI: 0.07, 0.23) had significantly decreased odds defaulting from the program.

Among those factors, only experiencing the staff giving out the wrong ration and the believing that the child was recovered overlapped with the model chosen during the original analysis with the incomplete dataset (Table 5). Both factors demonstrated similar trends of association though unequal magnitudes (aOR$_{original}$: 5.40; aOR$_{imputed}$: 27.87). Two factors that were included in the original model that did not get selected into the imputed model were feeling that SFP was too far away and believing that the child was not recovering.

In Kenya (Table 3), factors associated with an increased odds of defaulting were being busier this year compared to other years (aOR: 2.74; 95% CI: 1.28, 5.84), considering the SFP to be too far away (aOR: 4.08; 95% CI: 2.01, 8.27), and experiencing nomadic travel during the time of the program (aOR: 4.67; 95% CI: 2.18, 9.99). Other factors such as believing that the SFP should "weight [the] same village each day" (aOR: 9.54E-06; 95% CI: 7.44E-10, 0.12), having visited the SFP to find no food (aOR: 0.21; 95% CI: 0.08, 0.60), and feeling that the child was already recovered (aOR: 0.23; 95% CI: 0.10, 0.54) decreased the odds of defaulting from the program.

Compared to the original model, only one additional factor was selected which was believing that the SFP could be improved by "weight same village each day" (Table 6). This factor indicates that participants would find it useful if the entire village attended the clinic on the same day. Not included in the imputed model but chosen originally was believing that the way a patient was treated made them happy. All other factors cited above were also included in the original model wherein all effect trends remained constant with little shift in magnitude.

It was found that for the program in Sudan (Table 4), factors that decrease the odds of defaulting from the program were feeling that the child was not recovering (aOR: 0.42; 95% CI: 0.23, 0.75) and thinking that the child appeared recovered (aOR: 0.11; 95% CI: 0.07, 0.17). Conversely, if the child ever refused to eat the food (aOR: 2.08; 95% CI: 1.40, 3.09) and being too busy during the SFP (aOR: 3.12; 95% CI: 2.16, 4.51) significantly increased the odds of defaulting out of the program.

The imputed model included three factors that were also selected in the original model (Table 7): being too busy, feeling that the child was not recovering, and believing that the child was already recovered. The child in the program refusing to eat the food was the only additional factor included in the imputed model. Experiencing the child being ill during the program and citing that the SFP was too far away were dropped from the imputed model but included in the original model.

DISCUSSION

Continued developments have allowed data analysts to apply improved multiple imputation methods independent of the original researchers who implemented the study itself. This study applied these methods using SAS and fully conditional specific modeling strategies to categorical data that was focused around programmatic data of a supplemental feeding program in low-resource or emergency settings. Despite large percentages of missingness among some of the variables, the procedure successfully imputed all variables as intended. Past literature is mixed in its recommendations as to how many imputations should be chosen but the amount used in this study (n = 10) was sufficient for the level of missingness. Evident convergence during each step of the imputation indicates that 10 imputations were reasonable.

The models selected for each country confirmed and extended results from the models originally chosen during the initial analysis. For Chad, the original model did not include many factors likely due to them being dropped prior to analysis with a missing percentage over 15%. The overlapping factors were consistent in directionality. One factor in the imputed model produced an extremely high odds ratio leading the author to believe that the imbalanced responses led to an inflated output as opposed to an actual representation of the influence the factor has on a patient’s odds of defaulting.

In Kenya, all factors chosen in both models were consistent in their directionality. Furthermore, the precision of the measures was improved following imputation. The additional factor to the imputed model was significant with an extremely small adjusted odds ratio likely attributed to the low but non-zero frequency of positive responses. The two factors lacking in the imputed model compared to the original were found insignificant by a small margin in multivariate analysis.

The imputed model chosen for Sudan included two significant factors that were included in the original model but added two additional and missed three others. The odds ratios for the factors in common between the models were nearly identical in magnitude and precision. The additional factor was not noted in the original analysis meaning it was dropped from consideration due to its high percentage of missingness. A cutoff of 15% missing was applied to all factors before analysis began. The three factors lacking in the imputed model were found to be significantly individually-associated with the outcome and were dropped from consideration by a small margin. The expansion of the dataset may have impacted the significance to reflect its true non-association with the outcome when considered among all of the factors that would have been dropped in the original analysis.

Overall, the method worked surprisingly well. The majority of literature on MI focus on its application to longitudinal and continuous data. The addition of fully conditional specific methods has allowed the method to be applied to a wider range of variables under more complex scenarios.

Limitations

There were many computational obstacles in getting the procedure to run properly. In the end, it was found that the multivariate analysis would only run successfully on SAS version 9.3 and with java functions turned off. The reason for this was not found, nor addressed, among the discussions within the online SAS support community.

Though the methods have been developed far enough to be applied to categorical data, there is reason to suppose that it would be unsuitable for the present data. The theory of the method implies that the data surrounding the missing data will help to predict what a plausible value would be. This may not be true for programmatic data that have less to do with a patient's condition than their guardians' feelings towards the program.

CONCLUSIONS

Multiple imputation can serve well to produce an accurate conclusion by rectifying the effects of missing data within a large dataset. As seen in the comparison of the original and imputed models, it is clear that setting the cutoff at 15% missing for disqualifying factors in analysis may have led to missing significant factors in the conclusion. In this context, the conclusions of the original study are intended to directly inform programmatic decisions and intervention design. MI can be used to reaffirm the most important factors, found to be significant influencers in both the original and imputed studies. Future interventions should take both studies into consideration when designing supplemental feeding programs in emergency and low-resource settings, specifically those investigated here.

ACKNOWLEDGMENTS

I would not have succeeded in conducting this exercise without the encouragement of my advisor Dr. Kristin Wall, the assistance and patience of my colleague, and talented statistician, Erin Hulland, as well as the persistent example of Tess Palmer.

REFERENCES

[1] Berglund, P. and S. G. Heeringa (2014). Multiple imputation of missing data using SAS, SAS Institute.

[2] de Goeij, M. C., et al. (2013). "Multiple imputation: dealing with missing data." Nephrol Dial Transplant 28(10): 2415-2420.

[3] DiazOrdaz, K., et al. (2016). "Multiple imputation methods for bivariate outcomes in cluster randomised trials." Stat Med 35(20): 3482-3496.

[4] Little, R. J. and D. B. Rubin (2014). Statistical analysis with missing data, John Wiley & Sons.

[5] Messer, K. and L. Natarajan (2008). "Maximum likelihood, multiple imputation and regression calibration for measurement error adjustment." Stat Med 27(30): 6332-6350.

[6] Navarro-Colorado, C. (2007). "A retrospective study of emergency supplementary feeding programmes." Save the Children/ENN, juin.

[7] Navarro-Colorado, C., Shoham, J., and Mason, F. (2008). Measuring the Effectiveness of Supplementary Feeding Programmes in Emergencies, Humanitarian Policy Group: United Kingdom.

[8] Palmer, T. A., Bugli, D. G., Schroeder, M. S., Wall, K. M., Shoham, J., and Navarro-Colorado, C. (2017). Defaulting in Supplementary Feeding Programs: Post-Enrollment Risk Factors from Chad, Kenya, and Sudan.

[9] Pigott, T. D. (2001). "A review of methods for missing data." Educational research and evaluation 7(4): 353-383.

[10] Romaniuk, H., et al. (2014). "Multiple imputation in a longitudinal cohort study: a case study of sensitivity to imputation methods." Am J Epidemiol 180(9): 920-932.

[11] Rubin, D. B. (1996). "Multiple imputation after 18+ years." Journal of the American Statistical Association 91(434): 473-489.

[12] Schafer, J. L. (1997). Analysis of incomplete multivariate data, CRC press.

[13] Schroeder, M. S., Wall, K. M., Webb-Girard, A., Shoham, J., and Navarro-Colorado, C. (2016). Factors Affecting Defaulting in Children’s Supplemental Feeding Programs in Chad, Kenya, and Sudan.

[14] Soley-Bori, M. (2013). "Dealing with missing data: Key assumptions and methods for applied analysis." Boston University.

[15] Van Buuren, S., et al. (1999). "Multiple imputation of missing blood pressure covariates in survival analysis." Statistics in medicine 18(6): 681-694.

[16] Yuan, Y. C. (2010). "Multiple imputation for missing data: Concepts and new development (Version 9.0)." SAS Institute Inc, Rockville, MD 49: 1-11.

TABLES

Table 1. Subset of missingness patterns* analysis showing the frequency (*n*) and percentage (%) of occurrence by each unique pattern within the entire original dataset prior to imputation.

Group	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	*n*	%
1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	682	38.06
19	X	X	X	X	.	.	.	.	.	.	.	.	.	.	.	.	.	X	X	X	X	X	X	X	X	X	X	X	X	X	603	33.65
70	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	301	16.8
26	X	X	X	.	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	34	1.9
59	.	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	24	1.34
50	X	.	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	19	1.06
6	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	.	X	X	X	X	X	X	12	0.67
38	X	X	.	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	10	0.56
54	X	.	.	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	10	0.56
9	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	.	X	X	X	X	X	X	X	7	0.39
21	X	X	X	X	.	.	.	.	.	.	.	.	.	.	.	.	.	X	X	X	X	X	.	X	X	X	X	X	X	X	5	0.28
34	X	X	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	X	X	X	X	X	X	X	X	X	X	X	X	X	5	0.28
61	.	X	X	X	.	.	.	.	.	.	.	.	.	.	.	.	.	X	X	X	X	X	X	X	X	X	X	X	X	X	4	0.22
13	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	.	X	X	X	X	X	X	X	X	X	3	0.17
16	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	.	.	X	X	X	X	X	X	X	X	3	0.17
17	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	.	X	X	X	X	X	X	X	X	X	X	3	0.17
20	X	X	X	X	.	.	.	.	.	.	.	.	.	.	.	.	.	X	X	X	X	X	X	.	X	X	X	X	X	X	3	0.17
44	X	X	.	.	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	3	0.17
56	X	.	.	.	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	3	0.17

*****Not all variables are shown in this figure. Only those that are missing data were included to allow the table to fit on the page. Consequently, patterns such as group 70 are not completely missing but instead have an additional 30 variables that are fully complete. Additionally, variable names have been replaced by letters in consideration of space.

Table 2. Descriptive, Bivariate, and Two-Step Multivariate Analysis of Significant Factors Associated with Defaulting in Supplementary Feeding Programs (SFP) for Children 6 - 59 months (N = 6870) in **Chad** in 2010

Factor	Defaulters (N = 3540) (%)	Non-Defaulters (N = 3330) (%)	Bivariate Analysis				Initial Multivariate Analysis				Primary Multivariate Analysis			
			OR	95% Confidence Interval		P-value	OR	95% Confidence Interval		P-value	OR	95% Confidence Interval		P-value
				Lower	Upper			Lower	Upper			Lower	Upper	
Had problems getting to the SFP	464 (13.1)	449 (13.5)	1.016	0.505	2.043	0.9639								
How busy were you overall in the past 3 months?														
Less busy than other times	538 (15.2)	210 (6.3)	0.567	0.251	1.281	0.167	1.237	0.393	3.891	0.7152	1.259	0.459	3.457	0.6538
As busy as usual	2004 (56.6)	1341 (40.3)	*ref*				*ref*				*ref*			
More busy than usual	998 (28.2)	1779 (53.4)	2.666	1.625	4.373	0.0003	3.148	1.501	6.602	0.0026	3.208	1.673	6.151	0.0005
In relation to other years, was this:														
Busier than expected	681 (19.2)	1157 (34.7)	1.913	1.170	3.129	0.0109	0.910	0.447	1.852	0.795	0.837	0.434	1.613	0.5946
Less busy than expected	642 (18.1)	211 (6.3)	0.363	0.173	0.762	0.0085	0.325	0.121	0.869	0.0252	0.280	0.113	0.692	0.0059
As expected	2217 (62.6)	1962 (58.9)	*ref*				*ref*				*ref*			
How would you describe experience at SFP?														
Good	2587 (73.1)	1794 (53.9)	0.040	0.009	0.179	<.0001	0.028	0.004	0.186	0.0002	0.032	0.006	0.167	<.0001
Average	933 (26.4)	1191 (35.8)	0.074	0.017	0.331	0.0007	0.059	0.010	0.358	0.0021	0.074	0.015	0.369	0.0015
Bad	20 (0.6)	345 (10.4)	*ref*				*ref*				*ref*			
How could things be improved at SFP?														
Better staff training	3504 (99.0)	3313 (99.5)	0.034	0.000	27889	0.5906								
Provide shade in waiting area	2832 (80.0)	2497 (75.0)	1.019	0.137	7.599	0.9836								
Shorter waiting times	2083 (58.8)	2253 (67.7)	0.674	0.386	1.176	0.1558								

Factor	Defaulters (N = 3540) (%)	Non-Defaulters (N = 3330) (%)	Bivariate Analysis				Initial Multivariate Analysis				Primary Multivariate Analysis			
			OR	95% Confidence Interval		P-value	OR	95% Confidence Interval		P-value	OR	95% Confidence Interval		P-value
				Lower	Upper			Lower	Upper			Lower	Upper	
Give priority to cases from far	2826 (79.8)	2443 (73.4)	1.141	0.159	8.191	0.8847								
Attend new comers first	2603 (73.5)	2582 (77.5)	0.587	0.068	5.041	0.5928								
Ask to come less often	2994 (84.6)	2219 (66.6)	2.703	1.019	7.167	0.0462	3.650	1.421	9.380	0.0081	3.149	1.444	6.866	0.0045
Better quality food	1865 (52.7)	1352 (40.6)	1.659	0.707	3.891	0.2215								
Avoid days without food	3246 (91.7)	2840 (85.3)	1.510	0.212	10.762	0.6546								
Staff be more friendly	3097 (87.5)	2408 (72.3)	2.553	0.800	8.143	0.1042								
Be less strict with admission criteria	3125 (88.3)	3085 (92.6)	0.559	0.266	1.172	0.1208								
Open another SFP closer from home	3200 (90.4)	2846 (85.5)	1.350	0.428	4.255	0.5921								
Provide transport	2886 (81.5)	2123 (63.8)	2.500	1.342	4.661	0.0058	1.228	0.418	3.607	0.6974				
Weight same village each day	3446 (97.3)	3289 (98.8)	0.025	0.000	27759.406	0.5652								
Situation makes caretaker unhappy	1597 (45.1)	667 (20.0)	3.316	1.931	5.696	<.0001	1.714	0.804	3.655	0.1606				
Unhappy for Other Reasons	2235 (63.1)	1800 (54.1)	1.454	0.815	2.595	0.1923								
Other things were received from the SFP	1995 (56.4)	1937 (58.2)	0.924	0.490	1.743	0.7956								
Child Liked Food Received (CSB)	495 (14.0)	1306 (39.2)	0.253	0.127	0.504	0.0005	0.614	0.235	1.605	0.3097				

Factor	Defaulters (N = 3540) (%)	Non-Defaulters (N = 3330) (%)	Bivariate Analysis				Initial Multivariate Analysis				Primary Multivariate Analysis			
			OR	95% Confidence Interval		P-value	OR	95% Confidence Interval		P-value	OR	95% Confidence Interval		P-value
				Lower	Upper			Lower	Upper			Lower	Upper	
Child ever refused to eat the food	1330 (37.6)	991 (29.8)	1.425	0.928	2.187	0.1032								
Child Continued eating other foods as usual	936 (26.4)	820 (24.6)	1.103	0.735	1.655	0.6343								
SFP food was shared with others besides child	788 (22.3)	443 (13.3)	1.914	0.910	4.025	0.0835								
Did this aspect of the SFP make you happy?														
Time spent waiting in the centre	1848 (52.2)	1914 (57.5)	0.805	0.461	1.405	0.4268								
Comfort and shading of the waiting area	1387 (39.2)	1386 (41.6)	0.907	0.514	1.599	0.7223								
Staff competency	700 (19.8)	1067 (32.0)	0.527	0.284	0.979	0.0433	0.744	0.333	1.662	0.4631				
The type of food given (quantity or quality)	1938 (54.8)	2348 (70.5)	0.502	0.290	0.870	0.0163	1.108	0.503	2.438	0.7952				
The way your child was treated	1756 (49.6)	1982 (59.5)	0.667	0.386	1.151	0.1375								
The way you were treated	1473 (41.6)	1812 (54.4)	0.596	0.347	1.023	0.0596								
Did you experience any of the following during the time the child was following the nutrition programme?														
Experienced Illness of Child in the program	1440 (40.7)	910 (27.3)	0.548	0.495	0.607	<.0001	0.325	0.161	0.656	0.0018	0.362	0.191	0.688	0.0023
Illness of person normally accompanying the child	530 (15.0)	310 (9.3)	0.583	0.502	0.677	<.0001	1.550	0.586	4.101	0.377				

Factor	Defaulters (N = 3540) (%)	Non-Defaulters (N = 3330) (%)	Bivariate Analysis				Initial Multivariate Analysis				Primary Multivariate Analysis			
			OR	95% Confidence Interval		P-value	OR	95% Confidence Interval		P-value	OR	95% Confidence Interval		P-value
				Lower	Upper			Lower	Upper			Lower	Upper	
Mother pregnant or giving birth	370 (10.5)	150 (4.5)	0.404	0.332	0.492	<.0001	0.366	0.131	1.021	0.0549				
Illness of other family member	250 (7.1)	160 (4.8)	0.664	0.541	0.815	<.0001	3.224	0.960	10.831	0.0582				
Death in family/funeral	300 (8.5)	200 (6.0)	0.690	0.573	0.831	<.0001	1.441	0.408	5.088	0.5707				
Visiting Relatives	480 (13.6)	210 (6.3)	0.429	0.362	0.509	<.0001	1.246	0.460	3.372	0.6647				
No one to care for other children	70 (2.0)	100 (3.0)	1.535	1.126	2.091	0.0066	2.154	0.524	8.856	0.2869				
No one to accompany to SFP	40 (1.1)	40 (1.2)	1.064	0.684	1.654	0.7831								
Lost Card	150 (4.2)	70 (2.1)	0.485	0.364	0.647	<.0001	0.235	0.050	1.092	0.0647				
SFP too far	260 (7.3)	80 (2.4)	0.311	0.241	0.401	<.0001	0.602	0.160	2.273	0.4536				
Card withdrawn by SFP	640 (18.1)	60 (1.8)	0.083	0.064	0.109	<.0001	0.022	0.005	0.100	<.0001	0.030	0.009	0.100	<.0001
Told not to return by SFP staff	10 (0.3)	30 (0.9)	3.202	1.564	6.556	0.0015	15.823	0.784	319.399	0.0716				
Transferred to another program	0 (0.0)	20 (0.6)												
No food at SFP	0 (0.0)	10 (0.3)												
Didn't hear my name called out	10 (0.3)	10 (0.0)	1.063	0.442	2.559	0.891								
Staff were giving out incorrect ration	20 (0.6)	100 (3.00)	5.448	3.363	8.827	<.0001	31.52	6.435	154.5	<.0001	27.9	7.7	100.4	<.0001
Inconvenience of weighing day	0 (0.0)	0 (0.0)												
Unfriendliness of SFP staff	0 (0.0)	10 (0.3)												
Inconvenience of weighing day	0 (0.0)	0 (0.0)												

Factor	Defaulters (N = 3540) (%)	Non-Defaulters (N = 3330) (%)	Bivariate Analysis				Initial Multivariate Analysis				Primary Multivariate Analysis			
			OR	95% Confidence Interval		P-value	OR	95% Confidence Interval		P-value	OR	95% Confidence Interval		P-value
				Lower	Upper			Lower	Upper			Lower	Upper	
Unfriendliness of SFP staff	0 (0.0)	10 (0.3)												
Too busy	360 (10.2)	300 (9.0)	0.875	0.745	1.028	0.1041								
Nomadic travel	30 (0.9)	30 (0.9)	1.064	0.640	1.769	0.8119								
Labour migration	20 (0.6)	60 (1.8)	3.229	1.942	5.369	<.0001	4.289	0.316	58.13	0.2736				
No money for transport	70 (2.0)	40 (1.2)	0.603	0.407	0.892	0.0113	0.735	0.065	8.353	0.8034				
Costs associated with attending	10 (0.3)	0 (0.0)												
Involuntary displacement (fire, flood, outbreak)	50 (1.4)	70 (2.1)	1.499	1.039	2.161	0.0303	1.731	0.439	6.825	0.4322				
Festivity/ Marriage/Baptism	490 (13.8)	220 (6.6)	0.440	0.373	0.520	<.0001	1.165	0.468	2.897	0.7425				
Insecurity	10 (0.3)	0 (0.0)												
Child dislikes food	500 (14.1)	900 (27.0)	2.252	1.994	2.543	<.0001	0.815	0.305	2.177	0.6805				
Didn't feel the child was recovering	420 (11.9)	290 (8.7)	0.709	0.605	0.830	<.0001	0.256	0.098	0.671	0.0057	0.240	0.105	0.553	0.0008
Child seemed to be recovered	1950 (55.1)	480 (14.4)	0.137	0.122	0.154	<.0001	0.159	0.063	0.400	0.0002	0.123	0.065	0.231	<.0001
Husband/partner refused	50 (1.4)	30 (0.9)	0.635	0.402	1.000	0.0502								
Preferred traditional medicine	10 (0.3)	10 (0.3)	1.063	0.442	2.559	0.891								

Table 3. Descriptive, Bivariate, and Two-Step Multivariate Analysis of Significant Factors Associated with Defaulting in Supplementary Feeding Programs (SFP) for Children 6 - 59 months (N = 2970) in **Kenya** in 2010

Factor	Defaulters (N = 2210) (%)	Non-Defaulters (N = 760) (%)	Bivariate Analysis				Initial Multivariate Analysis				Primary Multivariate Analysis			
			OR	95% Confidence Interval		P-value	OR	95% Confidence Interval		P-value	OR	95% Confidence Interval		P-value
				Lower	Upper			Lower	Upper			Lower	Upper	
Had problems getting to the SFP	425 (19.2)	265 (34.9)	2.249	1.233	4.105	0.0083	1.670	0.581	4.802	0.3408				
How busy were you overall in the past 3 months?														
Less busy than other times	340 (15.4)	31 (4.1)	0.316	0.079	1.263	0.1028	0.353	0.055	2.261	0.2713				
As busy as usual	1413 (63.9)	394 (51.8)	*ref*				*ref*							
More busy than usual	457 (20.7)	335 (44.1)	2.630	1.469	4.708	0.0011	2.414	0.988	5.897	0.0532				
In relation to other years, was this:														
Busier than expected	843 (38.1)	450 (59.2)	2.367	1.279	4.381	0.0061	2.776	1.013	7.605	0.0471	2.739	1.284	5.842	0.0092
Less busy than expected	414 (18.7)	95 (12.5)	1.015	0.428	2.406	0.9724	5.089	1.234	20.992	0.0244	2.411	0.813	7.153	0.1126
As expected	953 (43.1)	215 (28.3)	*ref*				*ref*				*ref*			
How would you describe experience at SFP?														
Good	1241 (56.2)	302 (39.7)	0.188	0.039	0.895	0.0357	0.269	0.054	1.333	0.1077				
Average	938 (42.4)	418 (55.0)	0.344	0.073	1.619	0.1769	0.208	0.041	1.057	0.0583				
Bad	31 (1.4)	40 (5.3)	*ref*				*ref*							
How could things be improved at SFP?														
Better staff training	0 (0.0)	0 (0.0)												
Provide shade in waiting area	251 (11.4)	127 (16.7)	1.565	0.736	3.328	0.2446								
Shorter waiting times	643 (29.1)	259 (34.1)	1.260	0.714	2.221	0.425								
Give priority to cases from far	476 (21.5)	140 (18.4)	0.822	0.412	1.639	0.5768								

Factor	Defaulters (N = 2210) (%)	Non-Defaulters (N = 760) (%)	Bivariate Analysis				Initial Multivariate Analysis				Primary Multivariate Analysis			
			OR	95% Confidence Interval		P-value	OR	95% Confidence Interval		P-value	OR	95% Confidence Interval		P-value
				Lower	Upper			Lower	Upper			Lower	Upper	
Attend new comers first	154 (7.0)	50 (6.6)	0.941	0.329	2.690	0.9093								
Ask to come less often	83 (3.8)	35 (4.6)	1.178	0.246	5.637	0.8364								
Better quality food	1910 (86.4)	670 (88.2)	1.175	0.507	2.724	0.7065								
Avoid days without food	726 (32.9)	149 (19.6)	0.498	0.259	0.955	0.0359	1.747	0.560	5.446	0.3357				
Staff be more friendly	10 (0.5)	1 (0.1)	0.000	0.000	0.432	0.0367	0.000	0.000	0.444	0.0373	0.0001	2.1E-09	1.93	0.0635
Be less strict with admission criteria	382 (17.3)	165 (21.7)	1.326	0.682	2.581	0.4054								
Open another SFP closer from home	628 (28.4)	244 (32.1)	1.191	0.672	2.110	0.5489								
Provide transport	383 (17.3)	132 (17.4)	1.003	0.500	2.012	0.9943								
Weight same village each day	21 (1.0)	1 (0.1)	0.000	0.000	0.194	0.0278	0.000	0.000	0.340	0.0318	9.5E-06	7.4E-10	0.12	0.0215
Situation makes caretaker unhappy	625 (28.3)	238 (31.3)	1.156	0.647	2.067	0.6249								
Unhappy for Other Reasons	510 (23.1)	203 (26.7)	1.215	0.665	2.219	0.5271								
Other things were received from the SFP	2026 (91.7)	672 (88.4)	0.694	0.286	1.686	0.4195								
Child Liked Food Received (CSB)	1873 (84.8)	516 (67.9)	0.381	0.205	0.708	0.0023	0.379	0.129	1.112	0.0772				
Child ever refused to eat the food	1151 (52.1)	541 (71.2)	2.274	1.280	4.040	0.0051	1.187	0.429	3.281	0.7413				
Child Continued eating other foods as usual	1800 (81.5)	466 (61.3)	0.361	0.201	0.649	0.0007	0.968	0.379	2.470	0.9456				

Factor	Defaulters (N = 2210) (%)	Non-Defaulters (N = 760) (%)	Bivariate Analysis				Initial Multivariate Analysis				Primary Multivariate Analysis			
			OR	95% Confidence Interval		P-value	OR	95% Confidence Interval		P-value	OR	95% Confidence Interval		P-value
				Lower	Upper			Lower	Upper			Lower	Upper	
Child Continued eating other foods as usual	1800 (81.5)	466 (61.3)	0.361	0.201	0.649	0.0007	0.968	0.379	2.470	0.9456				
SFP food was shared with others besides child	2053 (92.9)	736 (96.8)	2.389	0.545	10.466	0.2478								
Did this aspect of the SFP make you happy?														
Time spent waiting in the centre	1059 (47.9)	286 (37.6)	0.656	0.383	1.124	0.1249								
Comfort and shading of the waiting area	1503 (68.01)	360 (47.4)	0.423	0.247	0.726	0.0018	1.279	0.484	3.376	0.6192				
Staff competency	1786 (80.8)	519 (68.3)	0.511	0.280	0.932	0.0286	0.755	0.240	2.370	0.6293				
The type of food given (quantity or quality)	867 (39.2)	202 (26.6)	0.561	0.312	1.006	0.0524								
The way your child was treated	1902 (86.1)	591 (77.8)	0.567	0.287	1.118	0.1012								
The way you were treated	2081 (94.2)	608 (80.0)	0.248	0.110	0.558	0.0007	0.254	0.063	1.020	0.0533				
Did you experience any of the following during the time the child was following the nutrition programme?														
Experienced Illness of Child in the program	830 (37.6)	200 (26.3)	0.594	0.495	0.713	<.0001	0.591	0.266	1.312	0.1964				
Illness of person normally accompanying the child	400 (18.1)	170 (22.4)	1.304	1.065	1.596	0.0101	1.426	0.498	4.081	0.5082				
Mother pregnant or giving birth	390 (17.7)	140 (18.4)	1.054	0.851	1.305	0.6309								

Factor	Defaulters (N = 2210) (%)	Non-Defaulters (N = 760) (%)	Bivariate Analysis				Initial Multivariate Analysis				Primary Multivariate Analysis			
			OR	95% Confidence Interval		P-value	OR	95% Confidence Interval		P-value	OR	95% Confidence Interval		P-value
				Lower	Upper			Lower	Upper			Lower	Upper	
Illness of other family member	180 (8.1)	60 (7.9)	0.967	0.713	1.311	0.8285								
Death in family/funeral	20 (0.9)	0 (0.0)												
Visiting Relatives	410 (18.6)	190 (25.0)	1.464	1.203	1.781	0.0001	0.833	0.325	2.137	0.7041				
No one to care for other children	280 (12.7)	90 (11.8)	0.926	0.719	1.193	0.5515								
No one to accompany to SFP	210 (9.5)	80 (10.5)	1.120	0.854	1.471	0.4124								
Lost Card	70 (3.2)	60 (7.9)	2.621	1.837	3.740	<.0001	4.434	0.609	32.29	0.1415				
SFP too far	450 (20.4)	400 (52.6)	4.346	3.644	5.183	<.0001	4.359	1.598	11.90	0.0041	4.075	2.008	8.268	<.0001
Card withdrawn by SFP	60 (2.7)	0 (0.0)												
Told not to return by SFP staff	30 (1.4)	0 (0.0)												
Transferred to another program	40 (1.8)	0 (0.0)												
No food at SFP	700 (31.7)	70 (9.2)	0.200	0.168	0.284	<.0001	0.100	0.031	1	0.0115	0.214	0.076	0.600	0.0034
Didn't hear my name called out	50 (2.3)	10 (1.3)	0.576	0.290	1.142	0.1142								
Staff were giving out incorrect ration	10 (0.5)	0 (0.0)												
Inconvenience of weighing day	430 (19.5)	210 (27.6)	1.581	1.306	1.913	<.0001	2.819	0.911	8.722	0.0721				
Unfriendliness of SFP staff	0 (0.0)	0 (0.0)												
Too busy	480 (21.7)	290 (38.2)	2.224	1.861	2.657	<.0001	1.632	0.625	4.261	0.3172				
Nomadic travel	250 (11.3)	300 (39.5)	5.113	4.202	6.222	<.0001	4.970	1.924	12.84	0.0009	4.665	2.178	9.992	<.0001
Labour migration	10 (0.5)	0 (0.0)												
No money for transport	60 (2.7)	50 (6.6)	2.525	1.718	3.711	<.0001	1.037	0.130	8.287	0.9725				

Factor	Defaulters (N = 2210) (%)	Non-Defaulters (N = 760) (%)	Bivariate Analysis				Initial Multivariate Analysis				Primary Multivariate Analysis			
			OR	95% Confidence Interval		P-value	OR	95% Confidence Interval		P-value	OR	95% Confidence Interval		P-value
				Lower	Upper			Lower	Upper			Lower	Upper	
Costs associated with attending	20 (0.9)	0 (0.0)												
Involuntary displacement (fire, flood, outbreak)	40 (1.8)	10 (1.3)	0.723	0.360	1.455	0.3635								
Festivity/ Marriage/Baptism	70 (3.2)	0 (0.0)												
Insecurity	40 (1.8)	30 (4.0)	2.231	1.379	3.610	0.0011	1.731	0.365	8.206	0.4895				
Child dislikes food	330 (14.9)	200 (26.3)	2.035	1.667	2.484	<.0001	2.572	0.879	7.527	0.0846				
Didn't feel the child was recovering	380 (17.2)	110 (14.5)	0.815	0.647	1.026	0.0819								
Child seemed to be recovered	640 (29.0)	90 (11.8)	0.330	0.260	0.418	<.0001	0.197	0.063	0.613	0.005	0.232	0.099	0.544	0.0008
Husband/partner refused	40 (1.8)	0 (0.0)												
Preferred traditional medicine	30 (1.4)	0 (0.0)												

Table 4. Univariate, Bivariate, and Two-Step Multivariate Analysis of Significant Factors Associated with Defaulting in Supplementary Feeding Programs for Households of Children 6 - 59 months (N = 6720) in **Sudan** in 2010

Factor	Defaulters (N = 4290) (%)	Non-Defaulters (N = 2430) (%)	Bivariate Analysis				Initial Multivariate Analysis				Primary Multivariate Analysis			
			OR	95% Confidence Interval		P-value	OR	95% Confidence Interval		P-value	OR	95% Confidence Interval		P-value
				Lower	Upper			Lower	Upper			Lower	Upper	
Had problems getting to the SFP	371 (8.1)	313 (9.0)	1.127	0.653	1.944	0.6675								
How busy were you overall in the past 3 months?														
Less busy than other times	3401 (73.8)	2491 (71.8)	0.800	0.513	1.246	0.323								
As busy as usual	504 (10.9)	462 (13.3)	*ref*											
More busy than usual	705 (15.3)	517 (14.9)	0.801	0.464	1.383	0.4252								
In relation to other years, was this:														
Busier than expected	1321 (28.7)	981 (28.3)	0.760	0.227	2.547	0.6561								
Less busy than expected	3224 (69.9)	2424 (69.9)	0.770	0.232	2.558	0.6687								
As expected	65 (1.4)	65 (1.9)	*ref*											
How would you describe experience at SFP?														
Good	3773 (81.8)	2904 (83.7)	1.184	0.452	3.101	0.7291								
Average	687 (14.9)	466 (13.4)	1.041	0.392	2.761	0.9355								
Bad	150 (3.3)	100 (2.9)	*ref*											
How could things be improved at SFP?														
Better staff training	108 (2.3)	199 (5.7)	2.561	1.055	6.213	0.0377	2.682	0.884	8.143	0.0815				
Provide shade in waiting area	3005 (65.2)	1858 (53.5)	0.438	0.135	1.419	0.1544								
Shorter waiting times	2097 (45.5)	1674 (48.2)	1.096	0.478	2.511	0.8162								
Give priority to cases from far	465 (10.1)	428 (12.3)	1.235	0.569	2.677	0.5796								
Attend new comers first	2241 (48.6)	1889 (54.4)	1.430	0.299	6.843	0.625								

Factor	Defaulters (N = 4290) (%)	Non-Defaulters (N = 2430) (%)	Bivariate Analysis				Initial Multivariate Analysis				Primary Multivariate Analysis			
			OR	95% Confidence Interval		P-value	OR	95% Confidence Interval		P-value	OR	95% Confidence Interval		P-value
				Lower	Upper			Lower	Upper			Lower	Upper	
Ask to come less often	0 (0.0)	0 (0.0)												
Better quality food	687 (14.9)	649 (18.7)	1.661	0.558	4.946	0.3563								
Avoid days without food	0 (0.0)	0 (0.0)												
Staff be more friendly	1219 (26.4)	740 (21.3)	0.681	0.272	1.702	0.3824								
Be less strict with admission criteria	449 (9.7)	641 (18.5)	5.891	0.846	41.031	0.071								
Open another SFP closer from home	21 (0.5)	36 (1.0)	24.15	0.000	5958563.6	0.5785								
Provide transport	40 (0.9)	46 (1.3)	1.649	0.126	21.615	0.6986								
Weight same village each day	0 (0.0)	0 (0.0)												
Situation makes caretaker unhappy	2290 (49.7)	1401 (40.4)	0.686	0.503	0.935	0.0171	1.326	0.822	2.138	0.2463				
Unhappy for Other Reasons	1391 (30.2)	646 (18.6)	0.529	0.376	0.746	0.0003	1.000	0.552	1.813	0.9998				
Other things were received from the SFP	4290 (100.0)	2420 (99.6)												
Child Liked Food Received (CSB)	4469 (96.9)	2948 (85.0)	0.176	0.082	0.374	<.0001	0.322	0.091	1.141	0.0778				
Child ever refused to eat the food	1170 (25.4)	1770 (51.0)	3.064	2.075	4.525	<.0001	2.066	1.269	3.362	0.0039	2.080	1.399	3.094	0.0004
Child Continued eating other foods as usual	3268 (70.9)	1823 (52.5)	0.455	0.297	0.695	0.0006	1.198	0.681	2.107	0.5269				
SFP food was shared with others besides child	1445 (31.3)	1039 (29.9)	0.934	0.641	1.360	0.7179								

Factor	Defaulters (N = 4290) (%)	Non-Defaulters (N = 2430) (%)	Bivariate Analysis				Initial Multivariate Analysis				Primary Multivariate Analysis			
			OR	95% Confidence Interval		P-value	OR	95% Confidence Interval		P-value	OR	95% Confidence Interval		P-value
				Lower	Upper			Lower	Upper			Lower	Upper	
Did this aspect of the SFP make you happy?														
Time spent waiting in the centre	3664 (79.5)	2662 (76.7)	0.853	0.567	1.282	0.4407								
Comfort and shading of the waiting area	3517 (76.3)	2693 (77.6)	1.077	0.767	1.512	0.6687								
Staff competency	3637 (78.9)	2785 (80.3)	1.091	0.723	1.648	0.6758								
The type of food given (quantity or quality)	3633 (78.8)	2844 (82.0)	1.222	0.848	1.761	0.282								
The way your child was treated	3715 (80.6)	2834 (81.7)	1.074	0.739	1.561	0.7073								
The way you were treated	3666 (79.5)	2651 (76.4)	0.839	0.512	1.374	0.4744								
Did you experience any of the following during the time the child was following the nutrition programme?														
Experienced Illness of Child in the program	3120 (67.7)	1740 (50.1)	0.480	0.439	0.526	<.0001	0.855	0.510	1.432	0.551				
Illness of person normally accompanying the child	660 (14.3)	450 (13.0)	0.892	0.784	1.014	0.0815								
Mother pregnant or giving birth	250 (5.4)	180 (5.2)	0.954	0.784	1.162	0.6415								
Illness of other family member	880 (19.1)	630 (18.2)	0.940	0.840	1.053	0.2882								
Death in family/funeral	590 (12.8)	280 (8.1)	0.598	0.515	0.695	<.0001	0.849	0.427	1.687	0.64				
Visiting Relatives	2210 (47.9)	1000 (28.8)	0.440	0.400	0.483	<.0001	1.007	0.540	1.880	0.982				

Factor	Defaulters (N = 4290) (%)	Non-Defaulters (N = 2430) (%)	Bivariate Analysis				Initial Multivariate Analysis				Primary Multivariate Analysis			
			OR	95% Confidence Interval		P-value	OR	95% Confidence Interval		P-value	OR	95% Confidence Interval		P-value
				Lower	Upper			Lower	Upper			Lower	Upper	
No one to care for other children	1130 (24.5)	480 (13.8)	0.495	0.440	0.556	<.0001	1.841	0.891	3.802	0.0991				
No one to accompany to SFP	1590 (34.5)	590 (17.0)	0.389	0.349	0.433	<.0001	0.509	0.250	1.038	0.0631				
Lost Card	30 (0.7)	80 (2.3)	3.601	2.361	5.491	<.0001	3.589	0.960	13.42	0.0576				
SFP too far	30 (0.7)	100 (2.9)	4.530	3.005	6.830	<.0001	4.420	0.999	19.56	0.0502				
Card withdrawn by SFP	0 (0.0)	0 (0.0)												
Told not to return by SFP staff	0 (0.0)	20 (0.6)												
Transferred to another program	0 (0.0)	20 (0.6)												
No food at SFP	130 (2.8)	60 (1.7)	0.6	0.445	0.826	0.0015	1.219	0.432	3.441	0.7084				
Didn't hear my name called out	0 (0.0)	0 (0.0)												
Staff were giving out incorrect ration	0 (0.0)	0 (0.0)												
Inconvenience of weighing day	0 (0.0)	10 (0.3)												
Unfriendliness of SFP staff	0 (0.0)	0 (0.0)												
Too busy	1190 (25.8)	1340 (38.6)	1.808	1.644	1.988	<.0001	3.399	2.083	5.547	<.0001	3.121	2.160	4.509	<.0001
Nomadic travel	90 (2.0)	170 (4.9)	2.587	1.996	3.354	<.0001	1.586	0.633	3.975	0.3253				
Labour migration	590 (12.8)	560 (16.1)	1.311	1.157	1.486	<.0001	0.876	0.484	1.585	0.661				
No money for transport	50 (1.1)	90 (2.6)	2.426	1.712	3.438	<.0001	1.871	0.513	6.818	0.3426				
Costs associated with attending	20 (0.4)	40 (1.2)	2.676	1.562	4.587	0.0003	1.024	0.094	11.10	0.9845				
Involuntary displacement (fire, flood, outbreak)	40 (0.9)	10 (0.3)	0.330	0.165	0.661	0.0018	0.331	0.038	3	0.3155				

Factor	Defaulters (N = 4290) (%)	Non-Defaulters (N = 2430) (%)	Bivariate Analysis				Initial Multivariate Analysis				Primary Multivariate Analysis			
			OR	95% Confidence Interval		P-value	OR	95% Confidence Interval		P-value	OR	95% Confidence Interval		P-value
				Lower	Upper			Lower	Upper			Lower	Upper	
Festivity/ Marriage/Baptism	800 (17.4)	290 (8.4)	0.434	0.377	0.501	<.0001	0.579	0.297	1.130	0.1091				
Insecurity	60 (1.3)	30 (0.9)	0.661	0.426	1.028	0.0659								
Child dislikes food	140 (3.0)	260 (7.5)	2.586	2.095	3.192	<.0001	0.565	0.139	2.300	0.425				
Didn't feel the child was recovering	310 (6.7)	500 (14.4)	2.335	2.011	2.711	<.0001	0.448	0.225	0.891	0.0221	0.416	0.231	0.752	0.0037
Child seemed to be recovered	3750 (81.3)	1360 (39.2)	0.148	0.134	0.163	<.0001	0.118	0.071	0.195	<.0001	0.110	0.071	0.169	<.0001
Husband/partner refused	10 (0.2)	20 (0.6)	2.667	1.246	5.705	0.0115	4.328	0.133	140.8	0.4096				
Preferred traditional medicine	0 (0.0)	60 (1.7)												

Table 5. Comparison of adjusted odds ratios of significant factors and their associated confidence intervals and p-values between the original model (missing data present) and the model determined from imputed datasets selected by logistic regression modeling strategies for **Chad**.

Factor		Original Model				Imputed Model			
		aOR	95% Confidence Interval		P-value	aOR	95% Confidence Interval		P-value
			Lower	Upper			Lower	Upper	
How busy were you overall in the past 3 months?	Less busy than other times					1.259	0.459	3.457	0.6538
	As busy as usual					*ref*			
	More busy than usual					3.208	1.673	6.151	0.0005
In relation to other years, was this:	Busier than expected					0.837	0.434	1.613	0.5946
	Less busy than expected					0.280	0.113	0.692	0.0059
	As expected					*ref*			
How would you describe experience at SFP?	Good					0.032	0.006	0.167	<.0001
	Average					0.074	0.015	0.369	0.0015
	Bad					*ref*			
How could things be improved at SFP?	Ask to come less often					3.149	1.444	6.866	0.0045
Did you experience any of the following during the time the child was following the nutrition programme?	Experienced illness of child in the program	0.33	0.22	0.51	<.0001	0.362	0.191	0.688	0.0023
	SFP too far away	0.319	0.123	0.831	0.0194				
	Card withdrawn from SFP					0.030	0.009	0.100	<.0001
	Staff were giving out incorrect ration	5.396	1.017	28.627	0.0477	27.867	7.736	100.381	<.0001
	Child dislikes food	1.898	1.166	3.091	0.0100				
	Didn’t feel the child was recovering					0.240	0.105	0.553	0.0008
	Child seemed to be recovered	0.156	0.103	0.237	<.0001	0.123	0.065	0.231	<.0001

Table 6. Comparison of adjusted odds ratios of significant factors and their associated confidence intervals and p-values between the original model (missing data present) and the model determined from imputed datasets selected by logistic regression modeling strategies for **Kenya**.

Factor		Original Model				Imputed Model			
		aOR	95% Confidence Interval		P-value	aOR	95% Confidence Interval		P-value
			Lower	Upper			Lower	Upper	
How busy were you overall in the past 3 months?	Less busy than other times	0.12	0.02	0.77	0.026				
	As busy as usual	*ref*							
	More busy than usual	2.05	0.81	5.19	0.128				
In relation to other years, was this:	Busier than expected	3.103	1.143	8.426	0.026	2.739	1.284	5.842	0.0092
	Less busy than expected	7.976	1.856	34.274	0.005	2.411	0.813	7.153	0.1126
	As expected	*ref*				*ref*			
How could things be improved at SFP?	Weight same village each day					9.54E-06	7.44E-10	0.12	0.0215
Did this aspect of the SFP make you happy?	The way you were treated	0.196	0.058	0.668	0.009				
Did you experience any of the following during the time the child was following the nutrition programme?	SFP too far away	6.594	2.547	17.072	0.0001	4.075	2.008	8.268	<.0001
	No food at SFP	0.077	0.02	0.3	0.0002	0.214	0.076	0.600	0.0034
	Nomadic travel	3.548	1.397	9.012	0.0078	4.665	2.178	9.992	<.0001
	Child seemed to be recovered	0.159	0.051	0.496	0.0016	0.231	0.099	0.542	0.001

Table 7. Comparison of adjusted odds ratios of significant factors and their associated confidence intervals and p-values between the original model (missing data present) and the model determined from imputed datasets selected by logistic regression modeling strategies for **Sudan**.

Factor		Original Model				Imputed Model			
		aOR	95% Confidence Interval		P-value	aOR	95% Confidence Interval		P-value
			Lower	Upper			Lower	Upper	
Did your child ever refuse to eat the food?						2.080	1.399	3.094	0.0004
Did you experience any of the following during the time the child was following the nutrition programme?	No one to care for other children	1.93	1.01	3.68	0.047				
	No one to accompany to SFP	0.49	0.25	0.93	0.030				
	SFP too far away	4.71	1.15	19.35	0.032				
	Too busy	3.335	2.29	4.92	<.0001	3.121	2.160	4.509	<.0001
	Didn't feel the child was recovering					0.416	0.231	0.752	0.0037
	Child seemed to be recovered	0.15	0.10	0.21	<.0001	0.110	0.071	0.169	<.0001

FIGURES

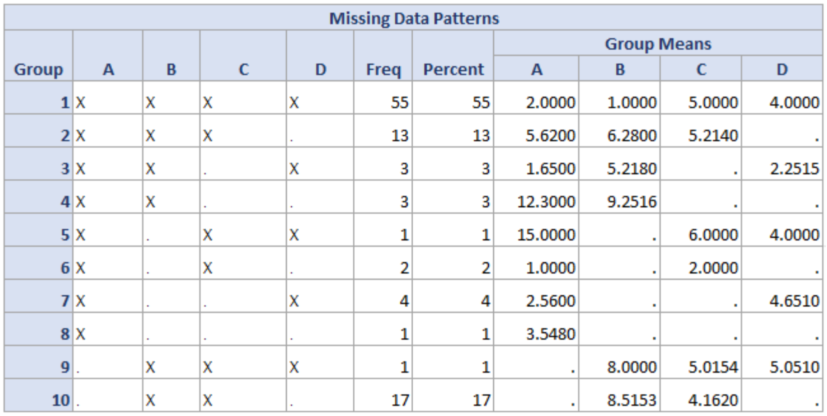
[1] Figure 1: Example of SAS (Version 9.4) code used to produce the frequency of missing items by variable.

```
PROC MEANS data=[dataset1] nmiss;
      VAR A B C D;
RUN;
```

[2] Figure 2: Example of SAS (Version 9.4) code used to produce the missing pattern figure as seen in Table 1.

```
PROC MI data=[dataset1] nimpute=0;
      VAR A B C D;
      ODS SELECT MISSPATTERN;
RUN;
```

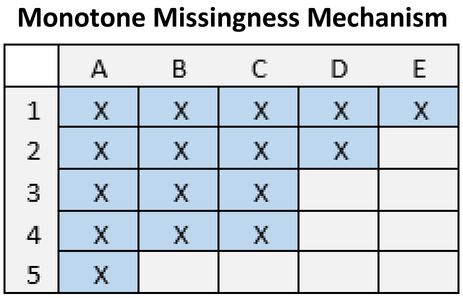
[3] Figure 3: Adapted output from SAS (Version 9.4) of the Missing Data Patterns produced by the PROC MI statement showing 10 unique patterns among four variables with their associated frequencies and group means.



Missing Data Patterns

							Group Means			
Group	A	B	C	D	Freq	Percent	A	B	C	D
1	X	X	X	X	55	55	2.0000	1.0000	5.0000	4.0000
2	X	X	X	.	13	13	5.6200	6.2800	5.2140	.
3	X	X	.	X	3	3	1.6500	5.2180	.	2.2515
4	X	X	.	.	3	3	12.3000	9.2516	.	.
5	X	.	X	X	1	1	15.0000	.	6.0000	4.0000
6	X	.	X	.	2	2	1.0000	.	2.0000	.
7	X	.	.	X	4	4	2.5600	.	.	4.6510
8	X	.	.	.	1	1	3.5480	.	.	.
9	.	X	X	X	1	1	.	8.0000	5.0154	5.0510
10	.	X	X	.	17	17	.	8.5153	4.1620	.

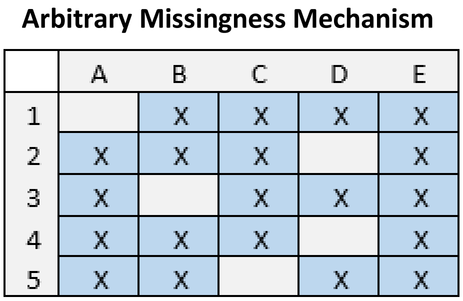
[4] Figure 4: Illustration of typical monotone missingness mechanism.



Monotone Missingness Mechanism

	A	B	C	D	E
1	X	X	X	X	X
2	X	X	X	X	
3	X	X	X		
4	X	X	X		
5	X				

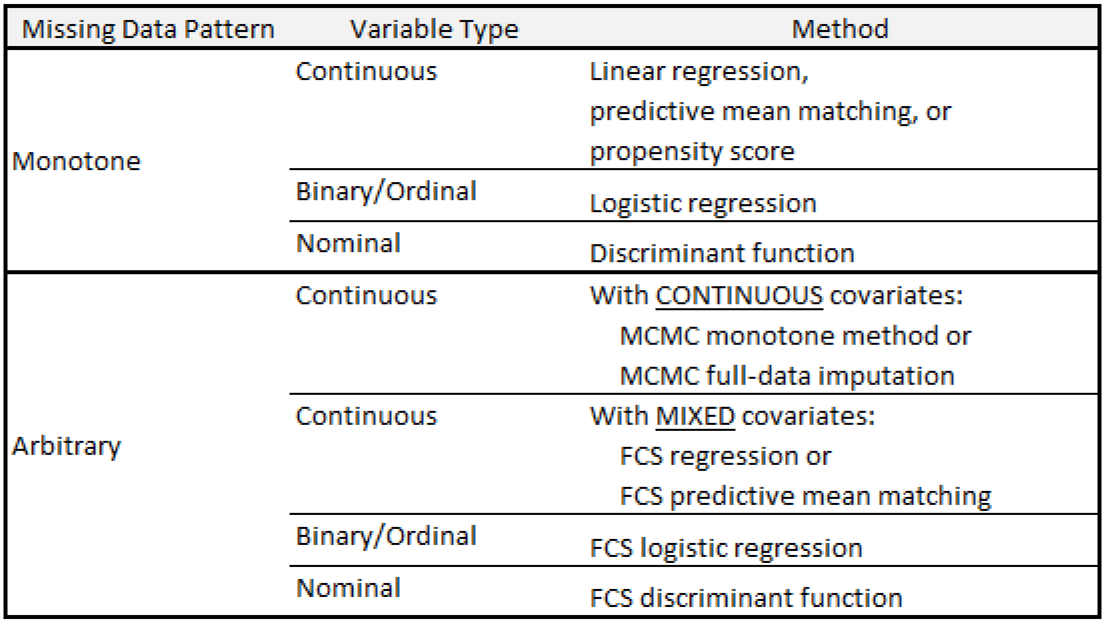
[5] Figure 5: Illustration of typical arbitrary missingness mechanism.



Arbitrary Missingness Mechanism

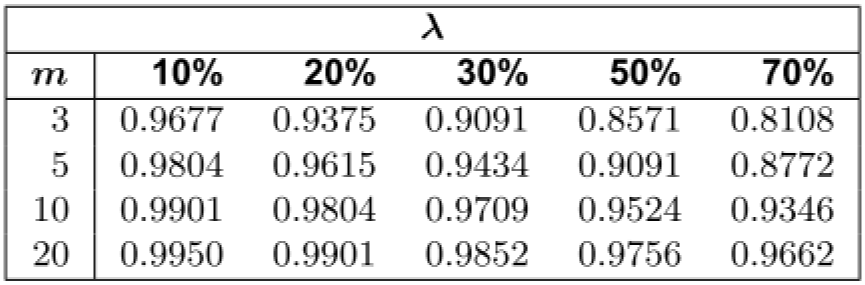
	A	B	C	D	E
1		X	X	X	X
2	X	X	X		X
3	X		X	X	X
4	X	X	X		X
5	X	X		X	X

[6] Figure 6: Imputation modeling method selection as determined by missing data pattern and variable type. Adapted from figure found in Berglund and Heeringa (2014).



Missing Data Pattern	Variable Type	Method
Monotone	Continuous	Linear regression, predictive mean matching, or propensity score
	Binary/Ordinal	Logistic regression
	Nominal	Discriminant function
Arbitrary	Continuous	With CONTINUOUS covariates: MCMC monotone method or MCMC full-data imputation
	Continuous	With MIXED covariates: FCS regression or FCS predictive mean matching
	Binary/Ordinal	FCS logistic regression
	Nominal	FCS discriminant function

[7] Figure 7: Multiple imputation efficiency by percentage missing as calculated by the formula proposed by Rubin (1987) and displayed in this table by Yuan (2010).



	λ				
m	10%	20%	30%	50%	70%
3	0.9677	0.9375	0.9091	0.8571	0.8108
5	0.9804	0.9615	0.9434	0.9091	0.8772
10	0.9901	0.9804	0.9709	0.9524	0.9346
20	0.9950	0.9901	0.9852	0.9756	0.9662

APPENDICES

APPENDIX I: A basic example of a PROC MI procedure

NOTE: There are many other options within the procedure that will not be mentioned, but this may serve as a reference for basic level multiple imputation.

```
PROC MI data=[dataset1] seed=12345 out=[dataset1]_mi
                        nimpute=10 nbiter=50
                        minimum=0  maximum=10;
      CLASS A B;
      VAR A B C D;
      FCS LOGISTIC (A = C D);
      FCS DISCRIM  (B = C D order=data classeffects=include);
RUN;
```

SAS CODE	SIGNIFICANCE
SEED=12345	A seed can ensure that you obtain the same imputations if the analysis ever needs to be repeated or validated.
NIMPUTE=10	The number of imputations can be manipulated with this option; the default is 5 imputations.
NBITER=50	This option allows for a custom burn-in iteration limit though the default, at 10, may not be reached if convergence is reached beforehand.

MINIMUM=0
MAXIMUM=10
	In dealing with continuous variables or categorical variables numerically represented, it may be useful to set a minimum and/or a maximum to avoid any illogical values as imputed values.

CLASS A B;	All categorical variables must be listed in the CLASS statement.
VAR A B C D;	All variables being used in imputation models must be listed within the VAR statement.
OUT=[DATASET1]_MI	Setting an OUT= statement allows the user to recall all of the imputed datasets as one large dataset which can be differentiated by the variable _imputation_ which is implicitly created by the PROC MI statement.
FCS	The FCS statement specifies a multivariate imputation by fully conditional specification methods. If you specify an FCS statement, you must also specify a VAR statement.
FCS LOGISTIC (A = C D);	This imputation model type is used primarily for the logistic regression of binary classification variables.
FCS DISCRIM (B = C D ORDER=DATA CLASSEFFECTS=INCLUDE);	Often used for logistic regression of classification variables with multiple levels, the discriminant function method requires an ORDER and CLASSEFFECTS option. ORDER specifies the order of the multiple levels while the CLASSEFFECTS statement ensures that the classification variables stated in the CLASS statement are included and treated appropriately.

SOURCE: https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#mi_toc.htm
https://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#statug_mi_sect008.htm

APPENDIX II: A basic example of a PROC SURVEYLOGISTIC procedure as a means of analyzing imputed datasets.

NOTE: There are many other options within the procedure that will not be mentioned, but this may serve as a reference for basic level multiple imputation. Any form of regression or survey methods can be used to analyze the data.

```
PROC SURVEYLOGISTIC data=x.das_mi;
      CLASS A (param = ref ref ='0');
      MODEL D (event = '1') = A;
      BY _imputation_;
      ODS output ParameterEstimates=out_mi;
RUN;
```

SAS CODE	SIGNIFICANCE
BY _IMPUTATION_;	The BY statement is necessary when analyzing the imputed datasets because this will create separate sets of parameter estimates for each unique imputed dataset.
ODS OUTPUT PARAMETERESTIMATES=OUT_MI;	Controlling the output of the procedure will vary depending on the method chosen, but it is important to obtain the parameter estimates and either their standard errors or covariances. These will be needed to run the subsequent procedure.

SOURCE: https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#surveylogistic_toc.htm

APPENDIX III: A basic example of a PROC MIANALYZE procedure as a means of combining parameter estimates across imputed datasets.

NOTE: There are many other options within the procedure that will not be mentioned, but this may serve as a reference for basic level multiple imputation.

```
PROC MIANALYZE parms(classvar=classval)= out_mi;
      CLASS A B;
      MODELEFFECTS A B C;
RUN;
```

SAS CODE	SIGNIFICANCE
PARMS(CLASSVAR=CLASSVAL)	Refers to a dataset that contains parameter estimates computed from the imputed data sets analyzed datasets in the analysis phase. If classification variables are included in the effects, the additional qualifier of CLASSVAR must be used with one of three options: FULL, LEVEL, or CLASSVAL.
CLASS A B;	CLASS statements are common throughout SAS methods and refer to the classification variables that are used as effects in the model being analyzed.
MODELEFFECTS A B C;	Similar to a VAR statement, the MODELEFFECTS statement will introduce the effects (individual variables or combined effects) to be used in the analysis. If an effect is stated in the MODELEFFECTS statement but not the CLASS statement, the procedure assumes it is continuous.

SOURCE: https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#mianalyze_toc.htm

APPENDIX IV: Sample piece of SAS code used for multiple imputation.

NOTE: This is a subset of the procedure applied to the data coming from Chad used in the present case study. The necessary CLASS and VAR statements are shown along with one example each of imputation models by FCS LOGISTIC and FCS DISCRIM methods.

```
PROC MI data=das_6 seed=1001 out=mi_das_c nimpute=10 minimum = 0;

      WHERE country = "C";
      CLASS x206_d    x211      x212_d    x213_d    x214_1_d x214_2_d
            x214_3_d  x214_4_d  x214_5_d  x214_6_d  x214_7_d x214_8_d
            x214_9_d  x214_10_d x214_11_d x214_12_d x214_13_d x215_d
            x216_d    x217_YN   x218_d    x219_num  x220_num  x222_d
            x301_HNH  x302_HNH  x303_HNH  x304_HNH  x305_HNH x306_HNH
            country;
      VAR   x206_d    x211      x212_d    x213_d    x214_1_d x214_2_d
            x214_3_d  x214_4_d  x214_5_d  x214_6_d  x214_7_d x214_8_d
            x214_9_d  x214_10_d x214_11_d x214_12_d x214_13_d x215_d
            x216_d    x217_YN   x218_d    x219_num  x220_num  x222_d
            x301_HNH  x302_HNH  x303_HNH  x304_HNH  x305_HNH x306_HNH
            x401_1_d  x401_2_d  x401_3_d  x401_4_d  x401_5_d x401_6_d
            x401_7_d  x401_8_d  x401_9_d  x401_10_d x401_11_d
            x401_12_d x401_15_d x401_16_d x401_19_d x401_20_d
            x401_21_d x401_22_d x401_23_d x401_24_d x401_25_d
            x401_27_d x401_28_d x401_29_d x401_30_d x401_31_d
            out2      sex       country;
      FCS LOGISTIC( x206_d =
                      x211      x212_d    x213_d    x214_1_d x214_2_d
            x214_3_d  x214_4_d  x214_5_d  x214_6_d  x214_7_d x214_8_d
            x214_9_d  x214_10_d x214_11_d x214_12_d x214_13_d x215_d
            x216_d    x217_YN  x218_d    x219_num  x220_num  x222_d
            x301_HNH x302_HNH  x303_HNH  x304_HNH  x305_HNH  x306_HNH
            x401_1_d  x401_2_d  x401_3_d  x401_4_d  x401_5_d x401_6_d
            x401_7_d  x401_8_d  x401_9_d  x401_10_d x401_11_d
            x401_12_d x401_15_d x401_16_d x401_19_d x401_20_d
            x401_21_d x401_22_d x401_23_d x401_24_d x401_25_d
            x401_27_d x401_28_d x401_29_d x401_30_d x401_31_d out2
            sex       country);
      FCS DISCRIM ( x211   =
            x206_d              x212_d    x213_d    x214_1_d x214_2_d
            x214_3_d  x214_4_d  x214_5_d  x214_6_d  x214_7_d x214_8_d
            x214_9_d  x214_10_d x214_11_d x214_12_d x214_13_d x215_d
            x216_d    x217_YN   x218_d    x219_num  x220_num  x222_d
            x301_HNH  x302_HNH  x303_HNH  x304_HNH  x305_HNH x306_HNH
            x401_1_d  x401_2_d  x401_3_d  x401_4_d  x401_5_d x401_6_d
            x401_7_d  x401_8_d  x401_9_d  x401_10_d x401_11_d
            x401_12_d x401_15_d x401_16_d x401_19_d x401_20_d
            x401_21_d x401_22_d x401_23_d x401_24_d x401_25_d
            x401_27_d x401_28_d x401_29_d x401_30_d x401_31_d out2
            sex       country / order=data classeffects=include);
RUN;
```