Multiple Imputation and Random Survival Forests: Application to the Demographic and Health Survey Child Survival Data

Afra Nuwasiima

Submitted to the School of Mathematics in partial fulfillment for a degree of Master of Science in Biometry

August, 2018

University of Nairobi

School of Mathematics

ISSN: 2410-1397

Master Project in Mathematics

Multiple Imputation and Random Survival Forests: Application to the Demographic and Health Survey Child Survival Data

Research Report in Mathematics, Number 08, 2018

Afra Nuwasiima

August 2018

Submitted to the School of Mathematics in partial fulfilment for a degree in Master of Science in Biometry

Master Project in Mathematics **University of Nairobi**
August 2018

Multiple Imputation and Random Survival Forests: Application to the Demographic and Health Survey Child Survival Data

Research Report in Mathematics, Number 08, 2018

Afra Nuwasiima

School of Mathematics
College of Biological and Physical sciences
Chiromo, off Riverside Drive
30197-00100 Nairobi, Kenya

Master of Science Project
Submitted to the School of Mathematics in partial fulfilment for a degree in Master of Science in Biometry

Prepared for The Director
Graduate School
University of Nairobi

Monitored by School of Mathematics

Abstract

Background: Demographic and Health Surveys (DHS) provide data on a wide scope of risk-factors of under-five child survival. Missing covariate data is inevitable in the DHS under-five survival data since data is collected retrospectively and on a large number of covariates. We studied the missing data problem on the risk-factors of under-five child survival in DHS data sets.

Methods: Random survival forests model was first used for selecting the highly predictive risk factors from a pool of over 400 covariates, from which a subset of 50 covariates was selected. Multiple imputation by chained equations (MICE) and random forests were applied to handle missing covariate data. Imputed data was then analyzed using random survival forests and Cox-regression models.

Results: The results showed that missingness in covariates was more related to the time to event (52%) than the event status (19%) response variables. The ranking of under-five risk factors from imputed data sets was closely related to the ranking from the observed values, albeit, multiple imputation led to increase in the variable importance scores. The unadjusted estimates from the Cox-regression model based on imputed values were closely similar to the estimates from the observed values. However, minimal discrepancies in estimates were observed in covariates with over 30% missing data. Random forests approach shown potential for producing estimates much closer to the true estimates with high level of missing than MICE.

Conclusion: Multiple imputation shown potential to produce estimates closely similar to the true estimates even with high missingness. Random forests imputation shown potential to perform better than MICE imputation strategies. The current study results may need to be validated using a larger simulation study and other non-response models for decisive conclusions to be made.

Dissertations in Mathematics at the University of Nairobi, Kenya.
ISSN 2410-1397: Research Report in Mathematics

DISTRIBUTOR: School of Mathematics, University of Nairobi, Kenya

Declaration and Approval

I the undersigned declare that this project report is my original work and to the best of my knowledge, it has not been submitted in support of an award of a degree in any other university or institution of learning.

Signature Date

Afra Nuwasiima
Reg No. I56/89630/2016

In my capacity as a supervisor of the candidate, I certify that this report has my approval for submission.

Signature Date

Dr Nelson Owuor Onyango
School of Mathematics,
University of Nairobi,
Box 30197, 00100 Nairobi, Kenya.
E-mail: onyango@uonbi.ac.ke

Dedication

This project is dedicated to my family for your continued support to me. You are the reason I keep going.

To the God Almighty, HIS blessings to me are uncountable.

Contents

Acknowledgments

I want to first thank God who has made this possible, it's because of HIM that I can face tomorrow.

I want to thank the DELTAS Africa Initiative SSACAB (Grant No. 107754/Z/15/Z) for funding my studies and stay in Kenya for a period of two years.

In a special way, I thank my supervisor and friend Dr. Nelson Owuor Onyango for his invaluable support towards my masters program. Merci beaucoup.

To you my friends and SST discussion group members at the school, I can't thank you enough, you made this masters lively.

Finally, to everyone who wished me well, especially my employer who allowed me ample study time, thanks a bunch.

Afra Nuwasiima

Nairobi, 2018.

1 Introduction

This chapter is divided into five subsections that include; background, statement of the problem, objectives, justification, and scope. The background includes an overview of child mortality, missing data, survival analysis, and classification and regression trees.

1.1 Child mortality

Child mortality estimated as rate of dying between 0 to 59 months is the principal measure for child well-being UNDESA (2015). Tremendous achievements have been made worldwide in reducing child deaths in the past two decades (Unicef, 2015). The figure 1 shows that the rate of child mortality has significantly decreased from 94/1000 deaths in 1990 to 41/1,000 in 2016, confirming a 55% reduction in child mortality in a period of 26 years. Globally, low and middle income countries (LMICs) account for about 99% of the under-five child deaths registered and sub-Saharan African (SSA) alone, accounts for about 50% of the under-five child deaths registered in the LMICs annually (Unicef, 2015).

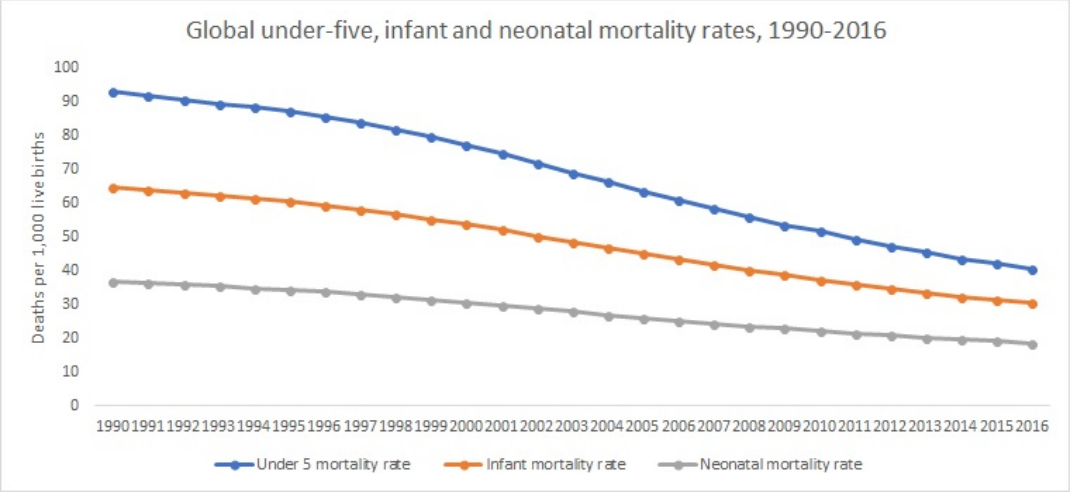


Figure 1. Global trends in under-five child mortality. Source: WHO estimates

Despite this progress, more efforts are still needed to realize the Sustainable Development Goal of preventing neonatal and child mortality and achieving below 25/1000 child deaths in every nation (Unicef, 2015). The figure 2 shows that diarrhea (29.6%), malaria (29.5%)

and AIDS (17.3%) are the most prevalent causes of neonatal mortality in the sub-Saharan African region by 2016 while Phenomena (22.5%), malaria (15.3%) and diarrhea (14.5%) were the most prevalent causes of post-neonatal mortality in the same region and year respectively. Albeit, 30.1% of the child deaths aged 1-59 months were accounted by other causes.

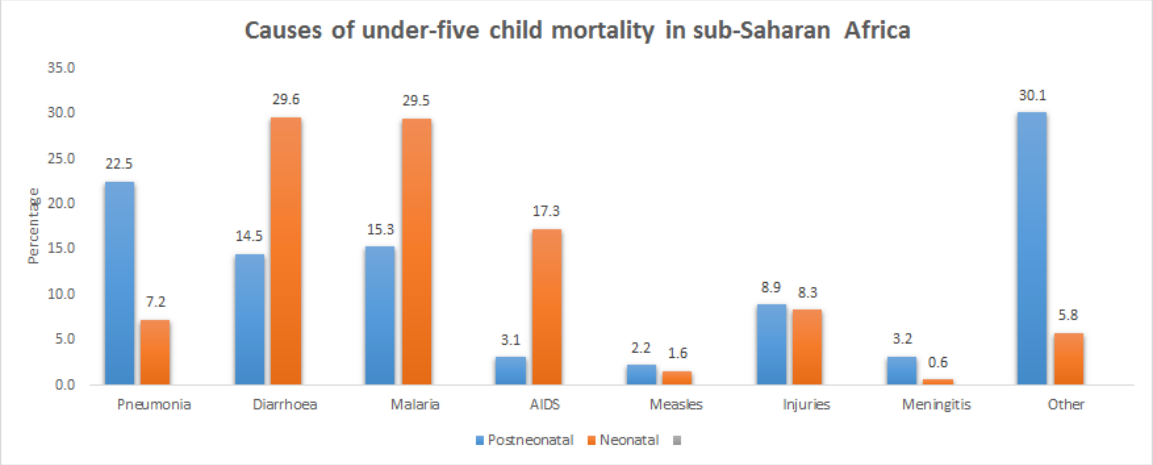


Figure 2. Causes of under-five child deaths in SSA. Source: UNICEF, 2016

Demographic and Health Surveys also collect data on a broad scope of risk factors of under-five child mortality. DHS are a series of national representative surveys that collect routine data on demographioc and child health indicators. DHS data sets provide a good source to understand some of the social, economic, demographic, enviromental, community and health risk-factors. Several studies Masanja et al. (2008); Susuman et al. (2016); Susuman and Hamisi (2012); Nasejje et al. (2015) have previously used DHS data sets to study the risk-factors of under-five child mortality in SSA. Several risk factors such as residing in rural areas, short preceding birth intervals, high parity, male children, high number of births and low mother's education were indicated as significant predictors of child survival. A study by (Nasejje et al., 2015) on under-five child survival using the Uganda DHS further identified high level of missing covariate data as one of the limitations to studying more important covariates. Several studies reviewed in this thesis did not report how they handled missing data, even when missing data was present in their data sets.

1.2 Missing data

1.2.1 Overview

Missing cases happens when there's no observed values for the given variable(s) or for the entire observation/unit. Missing data is inevitable in social and health sciences research (Allison, 2001). Missing values occur at two stages/levels of analysis namely; unit and

item stages (Azur et al., 2011). At unit level, missing data happens when data values are not recorded for a respondent for reasons such as the respondent is unavailable or refuses to complete the survey. At item level, missing data occurs when data is partially missing for the respondent i.e. data is recorded for some variables and not recorded for some variables. The reasons for this missing data may include refusal/forgetting to answer some of the questions, and skip patterns in the questionnaire. However, in longitudinal studies or randomized controlled trials, missing cases may occur due to loss to follow-up, withdrawal from the study, and death if not the study's interest (Allison, 2001)

Demographic and Health Surveys (DHS) experience both unit and item level missing data. Information from the 2015-2016 Tanzania Demographic and health Survey (TDHS) indicate that data was missing at unit level in 3% out of the total 13,634 eligible women identified. In this thesis, our focus is on item level missing data in the 2015-2016 TDHS. Data on item level is missing in over 60% of the data collected in the Demographic and Health Surveys. It's important to note that data is collected retrospectively for the previous five years in the DHS, and this plays a bigger role on the item level missingness. Data is also missing in DHS due to the skip patterns that are observed in the DHS Woman's questionnaire.

Missing data has also been reported in earlier reviews. (Peugh and Enders, 2004) assessed the prevalence of missing cases in education and psychology research. They found out that 48% of the articles reviewed contained missing data, and over 90% of the articles with missing data applied the conventional complete case or pair-wise deletion analysis methods. Several limitations of complete case analysis have been discussed in literature (Graham, 2009). If missing data is not handled properly it may lead to invalid statistical inferences (Graham, 2009). (Allison, 2001) indicated that the best way to deal with unobserved observations is to avoid them. Missing data may be avoided at the data collection stage by having strict data monitoring and collection teams that ensure complete records are achieved or follow up cases that record missing data. The use of electronic data collection systems like Open Data Kit (ODK) that are pre-programmed to limit missing information are handy in preventing the missing data problem.

There are other several alternatives of handling missing data that include imputation, weighting and maximum likelihood estimation techniques (Little and Rubin, 1989; Schafer and Graham, 2002). When handling missing observations, it is very paramount to pay attention to the following aspects;

- Proportion of missingness
- Pattern of missing data
- Missing data mechanisms

The three items above are described in our methods section.

1.2.2 Missing data imputation techniques

There are several techniques that are used in handling missing data. They include but not limited to;

Complete case analysis (CCA)

CCA refers to the analysis of only complete observations i.e. observations with missing cases do not form part of the analysis. When data is missing completely at random (MCAR) i.e when the probability of missing does not depend on the observed or unobserved, complete case analysis will be unbiased, otherwise, CCA will produce biased or unbiased estimates under the following scenarios (Little and Rubin, 2014);

- If the response variable (Y_i) is MAR or NMAR, complete case will be biased
- If the missingness in covariates (X_i) is dependent on the values of Y_i, the complete case will be biased.

Data is said to be MAR (Missing at random) if the probability of missing depends on the observed values where as data is said to MNAR (Missing not at random) if the probability of missing depends only on the unobserved values.

Single Imputation

First, imputation refers to taking draws or means from the posterior distribution of the observed values (Little and Rubin, 2014). Single imputation means that only a single draw is taken. There are two approaches of generating data under single imputation; explicit (direct) and implicit (indirect) modeling (Little and Rubin, 2014). Explicit modeling approaches include the use of means, regression, and stochastic regression. Implicit methods use hot-deck, substitution, and cold deck. Mean imputation is further broken down into conditional and unconditional mean imputation methods. We discuss the different single imputation approaches below;

Unconditional mean imputation

Let X_{ij} represent a value of X_j for observations $i = 1, 2, ..., n$. Let $\bar{x}$ represent the respondent data mean. The missing cases are estimated by the computed value $\bar{x}$. When the likelihood of missing is independent of observed or unobserved values, the variance estimates will be unbiased. Otherwise, estimates from this approach produce biased estimates (Little and Rubin, 2014).

Imputing means within adjustment cells

This approach is commonly used in surveys where observations are classified into B adjustment classes. The respondent mean is the same for the non-respondent mean if in one class, assuming equal sampling weighs. Let $\bar{x}_{jB}$ be the observed mean for a variable X_j in class B. Then the resultant mean of X from imputed data is

$$\frac{1}{m}\sum_{j=1}^{B}\left(\sum_{i=1}^{b_j} x_{ij} + \sum_{i=b_j+1}^{m_j} \bar{x}_{jB}\right) = \frac{1}{m}\sum_{j=1}^{B} m_j \bar{x}_{jB} = \bar{x}_{wc} \qquad \textbf{(1)}$$

Regression imputation (RI)

RI method replaces unobserved data with the predicted values from a regression of unobserved cases on observed cases (Little and Rubin, 2014). Let $X_1, X_2, ..., X_{m-1}$ be completely observed and X_m be observed for only the first k observations and unobserved for $n-k$ observations. RI estimates the regression of X_m on $X_1, X_2, ..., X_{m-1}$ based on k complete cases, and then fills the missing values with regression predictions. The missing values are imputed using the regression equation

$$\hat{x}_{im} = \hat{\beta}_{m.012....m-1} + \sum_{j=1}^{m-1} \bar{\beta}_{mj.12....m-1} x_{ij}, \qquad \textbf{(2)}$$

where $\hat{\beta}_{m.012....m-1}$ is the intercept and $\bar{\beta}_{mj.12....m-1}$ is the regression coefficient of X_j in the regression of X_m on $X_1, X_2, ..., X_{m-1}$ based on k observed values.

Stochastic regression imputation

Under this approach, missing observations are replaced by predictions from the regression line plus residuals drawn to account for the uncertainty in the predicted observations. Let

$X_1, X_2, ..., X_{m-1}$ be complete cases and X_m be observed only for the first k observations and unobserved for the last $n-k$ observations. Stochastic regression estimates the regression of X_m on $X_1, X_2, ..., X_{m-1}$ based on k complete values. Imputed values are predictions from the conditional draw.

$$\hat{x}_{im} = \hat{\beta}_{m.012....m-1} + \sum_{j=1}^{m-1} \bar{\beta}_{mj.12....m-1} x_{ij} + U_{im}, \quad \textbf{(3)}$$

Where U_{im} is the disturbance term with mean zero and variance $\hat{\sigma}_{mj.12....m-1}$

Hot deck imputation

In this approach, missing observations of a respondent are replaced by observations from the similar "donor" in the observed data (Andridge and Little, 2010). This approach is recommended for missing data imputation in cases where missingness is less than 10% and the data is either MAR or MCAR (Andridge and Little, 2010). Consider a sample h from a population of H units. If a out of H units are observed for a variable X, where $i = 1, 2, ..., h$ and $a < h$. The mean of X may be estimated as the mean of the observed and imputed data, written as follows;

$$\bar{x}_{HD} = \left\{ a\bar{x}_A + (h-a)\bar{x}^*_{MR} \right\} / h \quad \textbf{(4)}$$

where $\bar{x}_A$ is the mean of the respondent units, and

$$\bar{x}^*_{MR} = \sum_{i=1}^{a} \frac{Q_i x_i}{h-a}$$

where Q_i is the number of times x_i is as an imputed value of X, with $\sum_{i=1}^{m} Q_i = h - a$, the number of unobserved units.

Last observation carried forward

This technique is most applied in longitudinal studies where there are loss to follow up cases. It has been previously applied in medical studies (Pocok, 1983) but (Molenberghs et al,2004) reported that the approach produces biased estimates even under MCAR. Little and Rubin (2014) provides a description of the mathematical implementation of the approach. Suppose $y_i = y_{i1}, ..., y_{ih}$ is a $(h * 1)$ observed vector of outcomes for an individual i, with a possibility of having missing data. Let R_i denote the missing value indicator, with R_i=0 if not missing, and $R_i = 1$ if an individual drops out between observation times $(h-1)$ and h. Then data is observed for $y_{i1}, ..., y_{i,h-1}$, and missing for $(y_{ih}, ..., y_{iH})$. Therefore, for an individual i with missing values, missing values are taken as the last respondent observed/recorded value.

Multiple imputation (MI)

MI is a missing observation handling approach that handles missing data by substituting every missing case with multiple values (Rubin, 1976; Little and Rubin, 2014). MI approach provides valid statistical inferences, especially under MCAR and MAR missing data mechanisms, and overcomes the limitations that come with single imputation by generating multiple complete data sets that can be analyzed and results pooled to form valid inferences. MI is insensitive to violations of the non-normality assumption, which makes it further appealing. According to (Little and Rubin, 2014), m multiple imputations are drawn from the known distribution of X_{mis}. When using MI, missing cases are estimated based on Bayesian iterative simulation methods that are based on the posterior predictive distribution of the unobserved cases. Two main methods used for drawing imputed values under multiple imputation are;

- Data augmentation
- Gibb's sampling algorithm (Gibb's sampler)

The other approaches used to generate multiple imputations include;

- Resampling
- Random forests
- Monte-carlo simulation models

Data augmentation (DA)

DA is defined as the iterative approach of simulating the posterior distribution of θ based on the Expectation Maximization (EM) algorithm and multiple imputation. This method is further discussed by (Rubin, 1987; Little and Rubin, 2014). The equation (5) shows the posterior distribution for the model when data is either MAR or MCAR.

$$f(\theta|X_{obs}, M) \equiv f(\theta|X_{obs}) = constant \times f(\theta) \times f(X_{obs}|\theta) \quad \textbf{(5)}$$

where $fp(\theta)$ is a prior distribution and $f(X_{obs}\theta)$ is the distribution of the observed cases. According to (Little and Rubin, 2014), data augmentation is a modification of the Expectation Maximization (E-M)algorithm when the sample size is not large. The E step under EM algorithm corresponds to imputation (I), and M step corresponding to posterior (P).

First, start with initial values $\theta^{(0)}$ chosen from an approximation to the posterior density of θ. Let $\theta^{(t)}$ represent a value of θ drawn at iteration t;

I$^{\text{Step}}$: Draw $X_{mis}^{(t+1)}$ with the function $f(X_{mis}|X_{obs},\theta^{(t)})$
P$^{\text{Step}}$: Draw $\theta^{(t+1)}$ with the function $f(\theta|X_{obs},X_{mis}^{(t+1)})$

The above iterative procedure yields an imputed value from the joint posterior function of X_{mis},θ given X_{obs}. DA can be run m times to compute m independent and identically distributed imputations.

Gibbs sampling

The Gibbs sampler is named after a great physicist Josiah Willard Gibbs who introduced it and is further described by (Geman and Geman, 1987). It is defined as an iterative simulation approach that draws imputations from the joint distribution. The Gibbs sampling algorithm is described by (Little and Rubin, 2014) as follows. The Gibbs sampler generates draws from the distribution $f(x_1,.....,x_k)$ of k random variables. The initial values $x_1^{(0)},.....,x_k^{(0)}$ are drawn in someway from the predictive distribution of X_{mis}. For every new iteration $(t+1)$, imputed values are draws from the sequence of k conditional distributions given the values $x_1^{(t)},.....,x_km^{(t)}$ from the previous imputation iteration t ;

$$\begin{aligned}
x_1^{(t+1)} &\sim f(x_1|x_2^{(t)},x_3^{(t)},.....,x_k^{(t)}).\\
x_2^{(t+1)} &\sim f(x_2|x_1^{(t+1)},x_3^{(t)},.....,x_k^{(t)}).\\
x_3^{(t+1)} &\sim f(x_3|x_1^{(t+1)},x_2^{(t+1)},x_4^{(t)},.....,x_k^{(t)}).\\
&\vdots\\
x_k^{(t+1)} &\sim f(x_k|x_1^{(t+1)},x_2^{(t+1)},.....,x_{k-1}^{(t+1)})
\end{aligned} \tag{6}$$

The sequence $x^{(t)}=(x_1^{(t)},.....,x_k^{(t)})$ converges to an imputation from a joint function of $X_1,.....,X_k$. The Gibbs sampler can be done independently t and m times to generate t iterations and m imputations respectively from the joint function of θ and X_{mis}.

Expected-Maximization (EM) Procedure

The EM procedure is defined as an efficient process of computing the maximum likelihood estimates in the presence of unobserved values (Borman, 2004). The EM algorithm doesn't provide imputed data as is the case of multiple imputation. The EM algorithm derives its estimates directly by maximizing the likelihood function of the observed cases (Dong

and Peng, 2013). Each cycle of the procedure involves two steps i.e. the expectation (E) step followed by a maximization (M) step (Allison, 2012). Missing data is estimated in the E-step given the available values and the parameters are estimated by conditional expectation. The maximization of the likelihood function is done in the M-step using both the imputed and observed data (Dong and Peng, 2013). Suppose $\vec{X}$ is a random vector and θ is the unknown parameter to be estimated such that $f(\vec{X}|\theta)$ is maximized. The following likelihood function is used to estimate θ

$$L(\theta) = \ln f(\vec{X}|\theta) \tag{7}$$

Maximizing equation (7) gives the values of θ. Therefore, after the n^{th} iteration, θ is estimated by θ_n satisfying the inequality,

$$L(\theta) > L(\theta_n) \tag{8}$$

We also have to maximize the inequality;

$$L(\theta) - L(\theta_n) = \ln f(\vec{X}|\theta) - \ln f(\vec{X}|\theta_n) \tag{9}$$

Therefore, if $\theta = \theta_n$, then $L(\theta_n)$ and $L(\theta)$ will be equal. We select θ given that $L(\theta_n)$ is maximized under EM algorithm.

Full Information Maximum Likelihood Estimation (FIMLE)

FIMLE is a method of estimating missing values by maximizing the likelihood function for every missing case given the available cases (Allison, 2012). Contrary to multiple imputation, FIMLE doesn't not provide imputed data sets, it only estimates parameters directly using all the observations in the data sets by maximizing the likelihood function (Dong and Peng, 2013). FIMLE is also known as Direct Maximum Likelihood Estimation. The first step in FIMLE is construction of the likelihood function. Suppose we have n independent records $(i = 1, 2, ..., n)$ on m variables $(x_{i1}, x_{i2}, ..., x_{im})$.

$$L = \prod_{i=1}^{n} p_i(x_{i1}, x_{i2}, ..., x_{im}; \theta) \tag{10}$$

To get values of θ , we need to maximize the likelihood function L. Suppose the variables x_{i1}, x_{i2} are either MAR or MCAR. The joint probability for a given observation i is the probability of observing the rest of the variables $x_{i3}, ..., x_{im}$. Suppose the missing values are discrete, the joint probability is summed over all the possible values with missing data as follows;

$$f_i(x_{i3}, ..., x_{im}; \theta) = \sum_{x_{i1}} \sum_{x_{i2}} f_i(x_{i1}, x_{i2}, ..., x_{im}; \theta) \tag{11}$$

If the missing data are continuous, then we have

$$f_i(x_{i3}, ..., x_{im}; \theta) = \int_{x_{i1}} \int_{x_{i2}} f_i(x_{i1}, x_{i2}, ..., x_{im}; \theta)\, dx_{i2}\, dx_{i1} \quad \textbf{(12)}$$

The overall likelihood function is the product of the separate likelihood functions for all the observations regardless of missing or non-missing cases. Assuming that there are p observations with complete cases and $n-p$ observations with incomplete cases, the overall likelihood function becomes;

$$L = \prod_{i=1}^{p} f_i(x_{i1}, x_{i2}, ..., x_{im}; \theta) \prod_{i=p+1}^{n} f_i(x_{i3}, ..., x_{im}; \theta) \quad \textbf{(13)}$$

The likelihood function (13) is maximized to obtain parameter estimates θ. The limitation for FIMLE is the problem of factorization of the likelihood functions that is very complex practically.

(Bennett, 2001) provides a summary table of the different missing data approaches discussed and their associated effects on the statistical inference as shown in Table 1.

Table 1. Missing data approaches and associated effects. Adapted from (Bennett, 2001)

Approach	Bias	Variability
Complete/Available case (CC)	Biased if MAR/MNAR	Gives lower standard errors
Mean methods (MM)	Biased if MAR/MNAR	Gives lower standard errors
Least observation carried forward (LVCF)	Biased if MAR/MNAR	Underestimates variance but less than MM
Regression approach (RM)	Biased only if MNAR	Gives lower standard errors
Hot-deck method	Biased only if MNAR	Gives lower standard errors but less than MM, LVCF & RM
MI	Biased only if MNAR	Gives good estimates of standard errors
Markov Chain method	Unbiased for all mechanisms	Gives better estimates of standard errors
E-M procedure	Biased only if MNAR	Gives better estimates of standard errors
FIMLE	Biased only if MNAR	Gives accurate estimates of standard errors

1.3 Survival analysis

1.3.1 Definition

(Kleinbaum and Klein, 2010) defines survival analysis as a set of statistical analysis approaches where the outcome of interest is time to the event. Time may be in minutes,

hours, days, weeks, months or years recorded from the start of the study until when an event of interest occurs or until when the study period ends. Under survival analysis, time is usually referred to as **survival time**

An event may refer to the condition or experience of interest that is being studied such as a disease, recovery from the disease, relapse, remission, death, loss of a job,return to a job, marriage dissolution. There are studies where more than one event is considered at the same time in data analysis, the statistical problem is characterized as **recurrent events** or **competing risks problem**.

1.3.2 Common functions in survival analysis

Survival function.

This is the basic quantile employed to describe time to event observations. This is the probability of a subject surviving beyond time t.

$$S(t) = Pr(T > t) \tag{14}$$

For a continuous random variable t,

$$\begin{aligned} S(t) &= 1 - P(T \leq t) = 1 - F(t) \\ &= 1 - \int_t^{\infty} f(x)dx \\ \frac{dS(t)}{dt} &= -\frac{d\int_t^{\infty} f(x)dx}{dx} \\ -\frac{d(S(t)}{dt} &= f(t) \end{aligned}$$

Therefore

$$f(t) = -\frac{dS(t)}{dt} \Rightarrow S(t) = \int_t^{\infty} f(x)dx$$

Properties of $S(t)$

- $S(0) = 1$
- $S(+\infty) = 0$
- $S(t)$ is an increasing function of t

Hazard function

This is defined as the likelihood that an event occurs at a time t given a subject has

survived upto or beyond time t.

$$h(t) = \lim_{h \to 0} \frac{P(t \leq T < t+h | T > t)}{h}$$
$$h(t) = \lim_{h \to 0} \frac{P(t \leq T < t+h)}{hP(T > t)}$$
$$h(t) = \lim_{h \to 0} \frac{S(t) - S(t+h)}{hS(t)}$$
$$h(t) = \frac{f(t)}{S(t)}$$

since

$$f(t) = -\frac{dS(t)}{dt}$$
$$h(t) = -\frac{dS(t)}{dt} / S(t)$$
$$h(t) = -\frac{dln(S(t))}{dt}$$

Note that $h(t)$ is also an increasing function.

Cumulative hazard function

$$\begin{aligned} H(t) &= \int_0^t h(u)du \\ &= \int_0^t -\frac{dln(S(u))}{du} \\ &= -ln(S(t)) \end{aligned}$$

The mean residual life time is thus given by,

$$\begin{aligned} r(t) = E(T - t | T \geq t) & \\ &= \frac{\int_t^\infty (x+t) f(x) dx}{S(t)} \\ &= \frac{\int_t^\infty S(x) dx}{S(t)} \end{aligned}$$

The p^{th} percentile t_p is the solution of the equation $S(t_p) = 1 - p$

Censoring

A data point is said to be censored if the event time is unknown by the closure of the study period. Censoring may occur in survival settings if there is **loss to follow up**, **study ends when the event has not occurred**, or **withdrawal from the study**.

Let $T(T \geq 0)$ be a positive survival time random variable, $t(t \geq 0)$ be the non-negative specific value for T, δ (0-1) be the variable for censorship with 1 if events occurs and 0 if

no event, and $C(C \geq 0)$ be the positive fixed censoring variable. We define the different types of censoring as follows;

Right censoring. Right censoring occurs when a subject's exact event time becomes incomplete on the right side of the follow-up period due to loss to follow up or withdrawal or study ending before event happening. Let $C_1, ...C_n$ be $i.i.d$ random variables representing the censoring time associated with T. Due to censoring, we can only observe the pairs:

$$(X_1, \delta_1), ..., (X_n, \delta_n)$$

where

$$X_i \quad \text{is min}(T_i, C_i)$$

and

$$\delta_i = I(T_i \leq C_i)$$

where $\delta_i = \delta_1, \delta_2, ..., \delta_n$ contain the censoring information.

Left censoring. This occurs when the outcome occurs prior to the start of the study/enrollment. Due to this type of censoring, we only observe the pairs: $(X_1, \delta_1), ..., (X_n, \delta_n)$
where

$$X_i = \quad \text{Max}(T_i, C_i) = \begin{cases} T_i & \text{if} \quad T_i \leq C_i, \\ C_i & \text{if} \quad T_i > C_i \end{cases}$$

and

$$\delta_i = I(T_i \leq C_i) = \begin{cases} 1 & \text{if} \quad T_i \leq C_i, \\ 0 & \text{if} \quad T_i > C_i \end{cases}$$

Interval censoring. This occurs when event is observed between two observation times. Let the interval between the two periods be $(L_i, R_i) = (L_1, R_1), ..., (L_n, R_n)$, then if,

$$(L_i, R_i) = \begin{cases} (0, C_L) & ; \quad \text{then left censoring} \\ (C_T, \infty) & ; \quad \text{then right censoring} \end{cases}$$

Double censoring. This occurs in studies where there are two related events, one followed by the other e.g. disease progression where the onset of the disease is caused by a viral infection. In such a scenario, there are three variables of interest; time to infection, time between infection and the onset of the disease, and finally time to the onset of the disease (Sun, 2007). Due to this type of censoring,we only observe the pairs: $(X_1, \delta_1), ..., (X_n, \delta_n)$ for $i = 1, 2, ..., n$

$$X_i = Min(Max(T_i, L_i), R_i)$$

and

$$\delta_i = \begin{cases} 1 & \text{if} \quad X_i = T_i, \\ 0 & \text{if} \quad X_i = R_i, \\ -1 & \text{if} \quad X_i = L_i \end{cases}$$

Truncation. This occurs when only individuals who have met certain criteria are included in the study eg. a study of life styles of retirees in a community. Anybody who has not retired doesn't qualify to be in this study and are thus truncated. There are two forms of truncation;
Right truncation. This occurs when the everyone in the study has already registered the outcome of interest.
Left truncation.This occurs when every subject has experienced the outcome of interest before joining the study.

1.3.3 Parametric survival analysis models

These are survival models where the survival time is assumed to follow a given distribution. The following parametric distributions are commonly used in survival analysis;

- Exponential, weibull, exponential, log-logistic, generalized gamma

Functions of some of the parametric distributions are shown in table below;

Table 2. Functions of some of the parametric distributions

Distribution	$f(t)$	$S(t)$	$h(t)$
Exponential	$\lambda exp(-\lambda t)$	$exp(-\lambda t)$	λ
Weibull	$\lambda p t^{p-1} exp(-\lambda t^p)$	$exp(-\lambda t^p)$	$\lambda p t^{p-1}$
Log-logistic	$\frac{\lambda p t^{p-1}}{(1+\lambda t^p)^2}$	$\frac{1}{1+\lambda t^p}$	$\frac{\lambda p t^{p-1}}{1+\lambda t^p}$

1.3.4 Cox regression

Cox regression model Cox (1972) is a mathematical model used to analyze survival data. The model assumes the proportional hazards assumption i.e for a given predictor, the hazard ration is constant over time. It's of the form

$$\begin{aligned} h(t|X) &= h_0(t) \exp(\beta_1 X_1 + ... + \beta_k X_k) \\ &= h_0(t) \exp(\beta^t X) \end{aligned} \tag{15}$$

Where $X_1, ..., X_k$ are the predictors, $h_0(t)$ is the baseline hazard, and it's function is unspecified. This explains why Cox regression is referred to as a semi-parametric regression model. The Parameters βs are estimated using the maximum likelihood estimation technique. The Cox model is explored more in our methods section.

1.3.5 Parametric survival model vs. Cox regression model

Parametric survival model has the following characteristics;

- More consistent with theoretical survival function
- Very simplified
- Completeness i.e the known distribution for $h(t)$ & $S(t)$
- The survival time assumes an underlying distribution

Cox PH model has the following characteristics;

- Less consistent with theoretical $S(t)$
- The distribution of survival time is unknown
- The baseline $S(t)$ or $h(t)$ are not specified
- The survival time doesn't rely on assumed distribution
- The baseline is not necessary for estimation of the hazards ratio

1.4 Classification and regression trees (CART)

CART are a simple non-parametric procedure used for prediction and predictor selection (Breiman et al., 1984). Under CART, the predictors are recursively partitioned into a set of homogeneous classes such that there is homogeneity of observation in the same class with respect to the outcome.

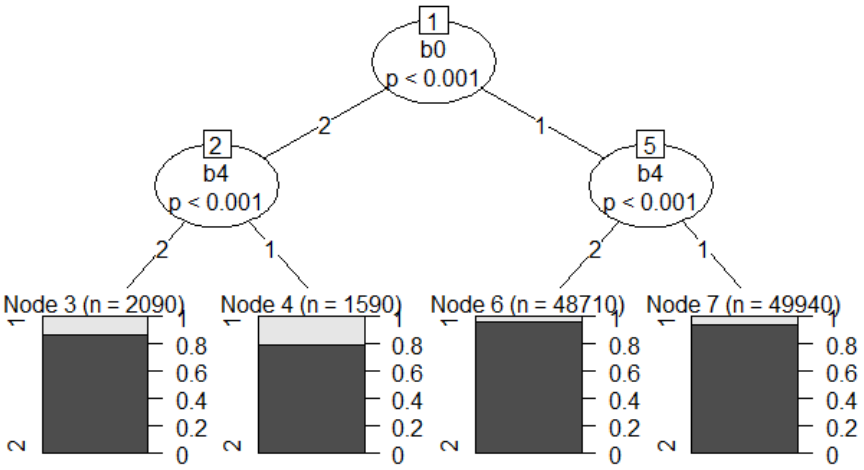


Figure 3. A pictorial of a conditional tree: The figure was constructed from the Tanzania DHS under-five child data. Observations in the final nodes are assumed to be homogeneous

1.4.1 Random Forests (RF)

RF is defined as a classifier that consists of a set of tree-structured classifiers

$$f(X, \Theta_m), m = 1, 2, ... \tag{16}$$

Where Θ_m are iid random variables and each tree produces a single vote for the class for every variable input X (Breiman, 2001). Random forests are an extension of the bagging where randomness is introduced (Breiman, 2001). RF splits the nodes using the best split point for a variable from a set of predictors randomly chosen at every node that provides more gainful information. RF fitting is based on the following algorithm (Breiman, 2001)

- Draw $ntree$ bootstrap samples from the original data. Reserve about 30% as test data (out of bag (OOB))
- Grow a tree for each $ntree$ bootstrap samples. At each node, use the best $m << p$ predictors and choose the best split among the m candidate predictors.
- For each bootstrap sample, predict the OOB data using the $ntree$ in the original sample
- Compute the error rate which is called the OOB estimate by averaging the OOB predictions.

1.4.2 Random Survival Forests

Random survival forests (RSF) are also ensemble tree methods designed for right censored observations (Ishwaran et al., 2008), and are an improvement of the random forests (Breiman, 2001). Both RF and RSF are part of classification and regression trees (CART) that are fully non-parametric. RSF were built to model complex interactions in the data apparent in survival data (Ishwaran et al., 2008) and have been used successfully in high dimensional cases where the covariates are greater than observations (Chen and Ishwaran, 2012). RSF tree nodes are built to maximize survival difference between two groups. RSF variable importance ranking procedure has been shown to be more stable than step-wise variable selection procedures and possesses high predictive performances. Unlike other statistical methods that have been used before to model the risk factors of child deaths, random survival forests have no restrictive assumptions that must be met before including the variables in the models. Fitting many survival trees enables random survival forests to estimate complex survival functions such as the non-proportional hazards with minimal prediction error Ehrlinger (2016). Hsich et al. (2011) and Hamidi et al. (2016) previously applied random survival forests for identification of important risk factors in systolic heart failure and kidney graft failure patients respectively. The mathematical description and implementation of RSF is discussed in our methods section.

1.5 Statement of the problem

Demographic and Health Surveys are a sequence of a national representative surveys conducted in LMICs to provide data on population, demographic and health measures. DHS data sets are good source of data on a wide scope of risk-factors of under-five child mortality. However, missing covariate data is inevitable in the under-five child survival DHS data sets since data is collected retrospectively and on a wide scope of risk-factors. Nasejje et al. (2015) studied factors of child survival in Uganda and identified high level of missingness on key covariates as a limitation to their study. Past studies either excluded variables with missing data or conducted a complete case analysis (CCA). CCA reduces the sample size and leads to reduced precison leading to underestimation of standard errors. MI is a flexible strategy of estimating missing covariate data that caters for the uncertainty about missing cases by creating many multiple imputed data sets, and pooling results from each data Rubin (1987). We examine the impact of imputing missing covariate data on the under-five child survival estimates using DHS data.

DHS data-sets are also composed of huge sets of variables and one is left with the challenge of choosing which variable to include in the study for analysis. This choice has to be statistical rather than by individual preference. Most studies have used litertaure reviews to select factors to include in the study. Howvever, the use of literature review may lead to exclusion of important covariates that have not been studied previously and is also biased towards individual's preference. Random Survival Forests (RSF) are a supplement of RF that are grown to handle right censored survival data while retaining all the appealing features of random forests Ishwaran et al. (2008). RSF have been shown to be sucessful in

identifying and ranking variables by using the inbuilt variable importance measures Chen and Ishwaran (2012). Unlike other statistical methods that have been used to model the risk factors of child mortality, random survival forests have no restrictive assumptions that must be met before including the variables in the models. The study applies random survival forests to rank variables by their order of importance.

1.6 Objectives

1.6.1 Overall objective

The study's main objective is to assess the effect of imputing missing covariate data on the under-five child survival estimates using the DHS data sets.

1.6.2 Specific objectives

1. To identify the highly predictive risk-factors of under-five child mortality from a pool of over 400 covariates
2. To assess the effect of imputing missing covariate data on the under-five child survival outcome
3. To compare the performance of different imputation strategies

1.7 Justification

High level of missingness in DHS data sets has constantly limited the scope of variables researchers can study Nasejje and Mwambi (2017). Identifying appropriate imputation strategies for DHS and alike data-sets will help users of such data-sets to consider studying a wide range of covariates. Proper handling of missing data in DHS data sets may also help increase on the validity of the conclusions drawn. The study applies further highly predictive models for variable selection on the imputed data to rank determinants of under-five child mortality from a pool of over 400 variables. The study uses a statistical approach to select variables as opposed to the contemporary use of literature that may leave out important predictors in the study.

1.8 Scope

The study uses the under-five child survival data set from the Republic of Tanzania. The study's population includes only under-five children born five years preceding the date of data collection (August 22, 2015, through February 14, 2016). This study covers a wide scope of covariates ranging from socio-demographic and -economic factors to health access behaviors.

2 Literature Review

2.1 Introduction

This chapter gives a brief outline of different subsections covered in this chapter. The chapter reviews first the factors affecting child mortality mentioning the methodological applications used. The study further reviews the different missing data handling techniques that have been applied previously.

2.2 Factors affecting under-five child mortality

Mosley and Chen (1984) designed an analytic frame work that has beenfor several years used for variable selection for child survival studies. The Mosley framework suggests that child survival needs to be investigated more as a chronic condition with multi-factorial origins rather than looking at it as acute-single cause phenomenon. The analytic framework groups the factors affecting child survival into two broad categories; proximate/immediate and socio-economic determinants.
The immediate factors include;

1. Maternal factors such as mother's age, gravidity, birth spacing
2. Environmental factors such as air/food/water/insects contamination
3. Feeding deficiencies i.e. inappropriate calories and micro-nutrients
4. Injury related causes i.e accidents
5. Personal health measures i.e prevention and control of illnesses

The socio-economic and demographic factors are are subdivided into three groups as follows;

- Individual level factors i.e parents,traditions/attitudes
- Household level factors i.e income
- Community level factors i.e health care systems, water source etc.

Susuman et al. (2016) studied the biological determinants of child survival in Tanzania. They employed binary logistic regression to model the determinats. The study's results shown that high mother's parity, short birth spacing, residing in rural areas significantly led to high child deaths. Four of the factors considered under this study contained missing data, and the researchers didnot mention how missing data was handled. Another study conducted in Tanzania Susuman and Hamisi (2012) on under-five mortality identified predictors such as poor mother's education,early age at first birth, short birth intervals, young mothers as significant contributors of under-five child survival in Tanzania. The study crude and adjusted odds ratios arising from the binary lostic model that was used. The study doesn't explain how variables were selected and the basis for variable categorizations. Armstrong Schellenberg et al. (2002) in a related study conducted in the 25 villages in the rural districts of Kilombero and Ulanga in the Southern Tanzania found out that high number of child deaths occurred due to fatal illnesses as a result of poor case management at the hospitals, low maternal education, nonexclusive breast feeding and lack of attendance of weighing clinics. Another interesting finding related to cultural practices was reported in this study. The study found out that there was a positive relationship between child deaths and mothers carrying babies on their back while cooking. The authors applied a binary logistic regression model to study the factors of child mortality.

Nasejje et al. (2015) conducted a study on understanding the predictors of under-five child survival in Uganda including accounting for the household and community shared random effects. This study found out that female headed households, male children and high number of baby deliveries in the past one year were significant predictors of under-five child deaths. The study used the Uganda Demographic and Health Survey data set and applied the frequentist and Bayesian approaches to study child survival. The study shown evidence of existence of unobserved shared effects at the household level but didn't find evidence to suggest existence of shared frailty effects at community level. This study indicated high level of missing covariate data as a limitation to the study to explore more important factors. The study further recommended the use of advanced models such as survival trees instead of the popular cox proportional hazards model that have restrictive assumptions.

Kozuki and Walker (2013) studied the link between long/short birth spacing periods and child loss, and found out that short intervals increaed the risk of child deaths compared to the middle interval. The study was conducted among the 47 low and income countries using logistic regression. Other studies conducted in Nigeria Abu et al. (2015); Ezeh et al. (2015); Yaya et al. (2017) reported earlier age at first sex, poverty, residing in rural areas, short preceding birth interval, lack of formal education and long distance to health facilities as significant factors associated with high under-five child deaths. These studies' findings are in sync with findings of similar studies reviewed above. The studies applied binary logistic and Cox-PH regression analysis methods to select important predictors.

2.3 Empirical review of missing data imputation approaches

Several studies have been conducted to evaluate the missing data approaches. Ma et al. (2011) conducted a study on the imputation approaches for unobserved binary responses in cluster Randomized Controlled Trials (RCTs) using a real and simulated data set. The study compared six different multiple imputation strategies that involved within- and across-cluster scenarios. The researchers found out that different imputation scenarios produce relatively similar findings with low missing values. However, the study found out that when data missingness is high, the imputation approaches that ignore the clustering underrated the standard errors of the treatment effect. The study concluded that strategies that catered for within-and across-cluster design are appropriate for RCTs.

Grund et al. (2017) studied the applicability of multilevel models when carrying out multiple imputations using simulated data sets. The study found out that multiple imputation with multilevel model structures provided less biased estimates than imputation models that ignored the multilevel structures. However, the study reported that multilevel multiple imputation with random slopes or interaction effects didn't yield reliable estimates, and recommended that future research should investigate multiple imputation using multilevel models that include random slopes and interaction effects.

Marshall et al. (2009) compared different imputation missing explanatory data strategies that included single imputation,and multiple imputation under different scenarios.The study fit a Cox regression model using a re-sampled data of 1000 observations while imposing different missingness mechanisms. The study reported that complete case analysis led to inefficient estimates when there were 25% or larger missing values. The study further found out that MI using predictive mean matching outperformed the other imputation strategies by producing the least biased parameter estimates. Another study Soullier et al. (2010) that examined the performance of MI under MCAR and MAR assumptions in estimation of occurrence rate in a cohort study found out that estimates under MCAR assumption were less biased compared to estimates computed under MAR. The confidence interval coverage rates under MCAR were also higher than those of MAR assumption.

Shah et al. (2014) conducted a study comparing random forest and the parametric imputation models for missing data imputation. The study reported no differences between predictive mean matching (PMM) and linear regression imputation models. The study found that random forests imputation models produced better estimates and confidence interval coverage values. The study recommended that random forests should be tested on larger and simulated data sets to assess it's performance in these settings. The study ignored clustering in the data and recommended that future research should consider using hierarchical models for imputation and analysis.

Jerez et al. (2010) also compared the performance of different statistical and machine learn-

ing approaches using a breast cancer data set. The study applied mean imputation,hot-deck and multiple imputation as statistical techniques and multi-layer perception,self-organization maps,and k-nearest neighbors as machine learning approaches. Multiple imputation was done using MICE and Amelia packages in R software. The study found that machine learning approaches outperformed other techniques in missing data imputation and improved significantly the accuracy of the predictions. Larsen (2011) compared full information maximum likelihood estimation (FIMLE) with second level dependencies and missing data imputation in a simulation study. The study reported that FIMLE produced less biased estimates when compared with multiple imputation. The study also indicated that when a more general imputation model than the analysis model is applied, better estimates were realized.

3 Data and statistical considerations

3.1 Data

In this thesis, we use data from the 2015-16 Tanzania DHS (TDHS), specifically the under-five children data set. The aim of the 2015-16 TDHS was to examine the patterns, levels, and trends of the health and population demographic measures. The survey employed a two stage sampling design. At stage one, a total of 608 clusters consisting of enumeration/community areas (EAs) drawn from the 2012 Tanzania Population and Housing Census were sampled. At stage two, a systematic sample of 22 households were sampled from each selected area. In total, 13,360 households were picked to participate in the study, and only 12,767 had occupants. Out of 12,767 occupied households, 12,563 were successfully enumerated. The data set provides information on every under-five child in the household including sex of the child, survival status of the child, birth interval, birth status, and child's weight at birth. The data also provides information on household and community characteristics, health coverage, maternal and antenatal care, infant feeding practices, and immunization coverage among others. The 2015-16 TDHS contains 10,233 observations and 1,253 variables. In this thesis, we define under-five child mortality as children who die between 1-59 months in the five years preceding the survey leaving us with 9,779 observations. The choice of 1-59 months was based on the need for the study to accommodate some of the survival analysis models that assume time (T) to be greater than zero (T>0). Our dependent variable was time to the event and event status. The event status was coded as 1=dead, 0=alive, and all the children alive were right censored.

Out of the total 1,253 variables, more than half them were excluded from the actual analysis data set using a procedure shown in Table 3.

Table 3. Data cleaning

Reason for excluding the variable	No. of variables excluded	Remaining no. of variables
Start	-	1253
100% missing	320	933
Index to birth history	8	925
Interview and sampling process information	38	887
Variables used to generate needed variables i.e dates, cmc	24	863
Variables recorded for only if child was alive	285	578
Flag variables/results of measurements	46	532
Assets variables used in the wealth index	26	506
Repeated/similar variables	98	408
Total considered		**408**

3.2 Statistical software and considerations

We used STATA version 13.0 College Station, TX: Stata Corp, USA for data cleaning including variable selection, and categorization. The cleaned data set was exported to R software version 3.4.1 and R Studio 1.1.153 R Foundation for Statistical Computing, Vienna, Austria for imputation and random forests implementation procedures.
All tests were considered statistically significant at 5% level of significance.

4 Random survival forests

In this thesis, we apply random survival forests (RSF) to select the highly predictive risk-factors of under-five child survival from the pool of 406 covariates considered in our thesis. RSF (Ishwaran et al., 2008) are highly predictive ensemble tree techniques applied to analyze right censored survival data to identify important risk-factors. RSF are an improvement of the random forests (Breiman, 2001) that fits right censored data. The following procedures were applied to rank the predictive risk-factors.

4.1 Random survival forests algorithm

Random survival forests was fit applied the algorithm introduced by (Ishwaran et al., 2008) shown below;

- Draw $ntree$ training samples from the original data. Every training sample excludes $\approx$ 37% of the cases as out of bag (OOB)/test data.
- Build a tree for every training data set. Randomly select m candidate variables ($mtry$) at every node of the tree ($m = \sqrt{p}$). The node is split using m variable that maximize survival difference between 2 daughter nodes. We applied log-rank splitting criterion as measure of survival difference.
- Build the tree to its full size under the restriction that terminal node contains utmost $nodesize = 3$ unique deaths.
- Calculate the cumulative hazard estimate (CHF) for every tree. The ensemble CHF is an average of the CHFs for all the $ntree$ trees.
- Compute an out-of-bag (OOB) prediction error rate for the tree using the OOB data

4.1.1 Log-rank splitting rule

Log-rank splitting rule plays a crucial role in the random survival forests algorithm by acting as a measure of node separation that helps in identifying the best split at a given node. The following log-rank splitting procedure (Ishwaran et al., 2008) was applied in this study for node splitting. Let v be a node of a tree, n be the individuals within node v. Let $(T_1, \sigma_1), ..., (T_n, \sigma_n)$ represent the survival times and censoring of the n individuals. Any split at node v on a given covariate x is of the form $x \leq s$ and $x > s$. The value of s is a random splitting value. Let $t_1 < t_2 < ... < t_N$ represent the individual survival times in

the node (v). Let $d_{i,j}$ represent event set and $Y_{i,j}$ represent the risk set at time t_i in the daughter nodes $j = 1, 2$. The log-rank test for a split at the value s for the predictor x is

$$L(x,s) = \frac{\sum_{i=1}^{N}(d_{i1} - Y_{i1}\frac{d_i}{Y_i})}{\sqrt{\sum_{i=1}^{N}\frac{Y_{i1}}{Y_i}(1-\frac{Y_{i1}}{Y_i})(\frac{Y_i-d_i}{Y_i-1})d_i}} \quad \textbf{(17)}$$

The best split at node v is decided by looking for the covariate x^* and split value s^* such that $|L(x^*, s^*)| \geq |L(x,s)| \quad \forall x$ and s. Other splitting rules are explained by (Ishwaran et al., 2008).

4.1.2 Ensemble estimation

Ensemble estimation is based on the cumulative hazard function (CHF) for every tree built in the forests. Consequently, ensemble CHF is obtained by averaging the CHF obtained on every tree for all the $ntrees$.
The CHF estimate for a node v is the Nelson-Aalen estimator shown below

$$\hat{H}_v(t) = \sum_{t_{l,v} \leq t} \frac{d_{l,v}}{Y_{l,v}} \quad \textbf{(18)}$$

If we get R terminal nodes in the tree, we will have R estimates of $\hat{H}_v(t)$. All the observations in v possess the same CHF. Every observation i contains a q-dimensional covariate x_i. Let $H(t|x_i)$ be the CHF for i. To estimate this value, drop x_i down the training tree. The CHF for i is the Nelson-Aalen estimator for $X_i's$ in node v

$$\hat{H}(t|X_i) = \hat{H}_v(t) \quad if \quad X_i \in v \quad \textbf{(19)}$$

Equation (19) defines the CHF for all the observations and the CHF for the tree. The CHF in equation (19) is computed for one tree. To compute an ensemble CHF, we average over $ntree$ trees in the forest. It should be noted that every tree in the forest is built using an independent training sample.

The bootstrap ensemble CHF for an observation i is

$$\hat{H}_e(t|X_i) = \frac{1}{ntree}\sum_{b=1}^{ntree}\hat{H}_b(t|X_i) \quad \textbf{(20)}$$

To compute the ensemble CHF for the OOB data, we let $I_{i,b} = 1$ if i is an OOB observation for $ntree$ training sample, and 0 if otherwise. The OOB ensemble CHF for the observation i is written as follows

$$\hat{H}_e^*(t|X_i) = \frac{\sum_{b=1}^{ntree} I_{i,b}\hat{H}_b^*(t|X_i)}{\sum_{b=1}^{ntree} I_{i,b}} \quad \textbf{(21)}$$

$\hat{H}_e^*(t|X_i)$ is an average over the training samples where i is an OOB observation. Note that Equation (20) uses all the $ntrees$ in the training sample unlike equation (21) which uses only the OOB/test observations.

4.1.3 Prediction error

The study estimated prediction error from the grown forest using the Harell's concordance index (C-index). The C-index estimates the likelihood that in a randomly chosen pair of cases, the case that fails first has a worst predicted outcome (WPO). To estimate C-index, we have to define what constitutes a WPO. Given a pair of cases (i, j), a case i will possess a WPO than j if

$$\sum_{k=1}^{N} \hat{H}_e^*(t_k|X_j) < \sum_{k=1}^{N} \hat{H}_e^*(t_k|X_i) \quad \textbf{(22)}$$

Now that we know how to estimate a WPO, we compute the C-index using the following rules;

1. Form all viable pairs of cases in the data
2. Omit those pairs whose shorter survival time (T) is censored. Additionally, omit pairs i and j if $T_i = T_j$ unless i or j is an event. Al the remaining pairs are defined as admissible pairs.
3. For each admissible pair where $T_i \neq T_j$, count 1 if the shorter T has a WPO. Count 0.5 if the predicted outcomes are tied. For each admissible pair where $T_i = T_j$ and both are events, count 1 if predicted outcomes are tied and count 0.5 if otherwise. For each admissible pair where $T_i = T_j$ but both are not events, count 1 if the event has a WPO and count and count 0.5 if otherwise.
   We define the concordance as the sum of the total counts for all the admissible pairs.
4. The concordance index (C-index) $= \frac{\text{Concordance}}{\text{Admissible}}$

Using the same procedure, compute the OOB estimate of C denoted as C^{**} using the OOB cases. The OOB prediction error (PE^{**}) is computed as $1 - C^{**}$. It is important know that $0 \leq PE^{**} \leq 1$. The prediction error rate of 0.5 equates to a random toss, and the error rate of 0 equates to perfect prediction.

4.1.4 Variable Importance

Best predictors for under-five child survival were selected based on variable importance (VarImp) measure (Ishwaran et al., 2008). VarImp can be interpreted in-terms of mis-classification. The VarImp for X quantifies the increase or decrease in the mis-classification

error on the OOB cases if X were unavailable. To compute VarImp for a risk-factor X, the OOB cases are dropped down their training survival tree. Whenever the split for X is met, the daughter node is assigned randomly. The new ensemble CHF is calculated, and the resulting prediction error is computed. The VarImp for X is then computed as the New Prediction error obtained after conducting random assignments subtracting the Original Prediction error. Large values of VarImp show risk factors with high predictive potential while zero or negative values of VarImp show predictors with no predictive ability.

4.1.5 RSF missing data imputation

Random survival forests have inbuilt measures of handling observations with missing data (Ishwaran et al., 2008) when splitting the variables at the parent node. The following steps are are applied only for observations with missing entries for the m candidate variables at node v.

1. For every node v, impute missing values before for splitting. Let $X^0_{r,v}$ represent the observed values for the r^{th} coordinate of the X-covariates in the training data in node h. Let $f(X^0_{r,v})$ be the posterior distribution of $X^0_{r,v}$
2. For every case in the training data in node v with missing data for the r^th coordinate, impute missing data by making draws from $f(X^0_{r,v})$. Redo this step for every r.
3. Split the node v using the imputed data applying the splitting rules described above.
4. After splitting the parent node v, reset the imputed values in the daughter node to missing.
5. Redo as in step 1 above for every node until the tree has reached it's saturation point.
6. The OOB cases are also imputed using the same rule.

4.2 Results

We fit the random survival model basing on the procedures above on the total of 408 variables. Figure 4 shows the variable ranking of the risk-factors that were considered in the data set.

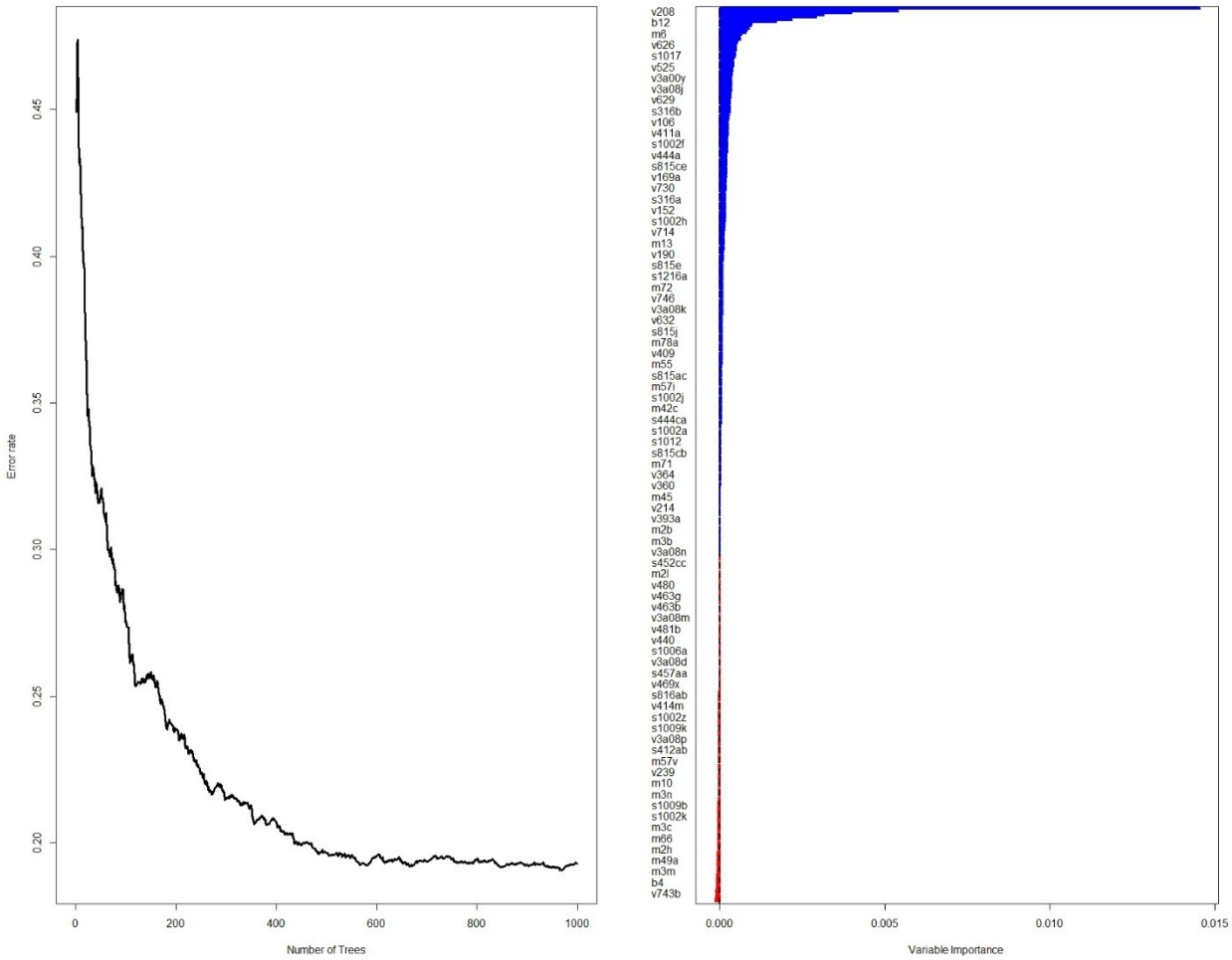


Figure 4. The left figure shows the prediction error rate and the right figure shows the the ranking of risk factors of under-five child mortality by order of importance. The prediction error rate was about 12%.

The table 4 below shows the variable importance scores for the first 10 highly ranked risk-factors. Based on (Ishwaran et al., 2008), all predictors with VarImp < 0.002 are less predictive. Table 4 shows that only 7 risk factors were highly predictive. Predictor b12 contains 70% missing data.

Table 4. VarImp scores in-order of high importance

Risk-factor	VarImp score
v137-total no. of under-five children	0.015
v208-no. of births in last one year	0.005
m4-child's breast feeding status	0.004
m5-months of breastfeeding	0.003
v238-births in last 3 years	0.003
v136-total no. of HH members	0.002
b12-Succeeding birth interval	0.002
v113-Source of drinking water	0.001
sreg1-region of residence	0.001
...	< 0.001

Our motivation was to assess the imputation approaches in a large data setting. Hence considered the first 50 highly ranked predictors for our next section.

5 Multiple Imputation (MI)

In this thesis, we applied MI techniques to impute unobserved covariate values. MI is a flexible missing data imputation strategy that handles missing data using three steps (Rubin, 1987) namely; imputation of multiple data sets, analysis of individual data sets separately and pooling results of multiple data sets together to come up with model estimates.

5.1 Proportion of missingness

The proportion of missingness plays an important role in deciding the choice of missing data approach. Prior reviews suggest that 5% or less missingness in the data may not affect the validity of the statistical inferences (Schafer, 1999). There's however, no clear consensus on the proportion of missing data that's statistically acceptable to produce valid inferences. (Bennett, 2001) suggests that if the data contains more than 10% missingness, the statistical inferences will be biased.

Out of the total 50 covariates considered, 23 (46%) covariates had missing cases. Average missing rate was 18% (0.01% to 92%). No specific code for skip patterns was created, hence covariates with missing data due to skips were treated as missing. Nonetheless, complete case analysis would produce unbiased estimates.

Variable type

Out of the total 23 variables with missing data, 5 were numeric, 4 were binary, 1 was ordinal($>$ 2 levels) and 13 were unordered ($>$ 2 levels).

5.2 Missing data patterns

Missing data pattern relates to the structure of the missing data matrix and doesn't relate to the association between the unobserved values and the available values. There are three missing data patterns that are discussed in literature (Dong and Peng, 2013) namely;

- Univariate missing data pattern
- Monotone missing data pattern
- Arbitrary missing data pattern

Let Y represent a data matrix with variables $Y_1, Y_2, ..., Y_m$. Data is said to possess a univariate missing data pattern if the same respondents have unobserved cases on one or more of the m-variables. The univariate missing data pattern will be prevalent in the data set if data is missing due to the design or skip patterns such that the same group of participants that don't meet the skip or design criteria will miss a given set of information.

A data-set containing variables $Y_1, Y_2, ..., Y_p$ is said to possess a monotone missing data pattern if in the event that data on a variable (Y_j) is unobserved for a particular respondent, all succeeding variables $(Y_{j+1}, Y_{j+2}, ..., Y_m)$ are as well partially unobserved. Such a missing data pattern occurs in longitudinal studies where whenever a subject drops out at a given time point, all future measurements for the same subject are unobserved.

A data set is said to possess an arbitrary missing data pattern, also known as general missing data pattern if the data is missing randomly for any given respondent or variables.

The choice of the imputation model must take into account the type of missing data pattern for valid imputations and consequent statistical inferences to be realized. For example if a data set possesses a monotone or univariate missing data pattern and the variables are continuous, imputation models that assume multivariate normality or that use propensity scores are appropriate namely regression, predictive mean matching and propensity scores (Schenker and Taylor, 1996). If the data are ordinal or nominal, logistic regression or discriminant function methods respectively are appropriate. However if data possess an arbitrary/general missing data pattern, Markov Chain Monte Carlo (MCMC) (Schafer and Olsen, 1998) or the chained equations are appropriate. Random forests imputation methods (Breiman, 2001; Doove et al., 2014) have the potential to model all the complexities and interactions and are appropriate for any type of missing data pattern. In this thesis, we focus more on both chained equations and random forests multiple imputation strategies.

5.2.1 Checking for missing data pattern

- Requires knowledge of the data. For-example, we don't expect monotone pattern in this study, not a longitudinal study
- There are several graphical tests (Templ et al., 2011) that are used to check for the pattern of missing data pattern.

Figures 5 and 6 show that our under-five child survival data possessed both univariate and arbitrary missing data patterns.

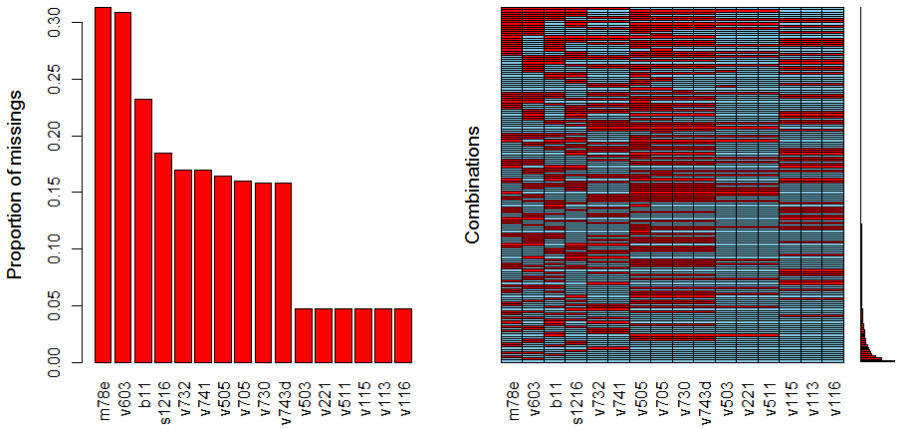


Figure 5. Missing data pattern plot. Shows the presence of univariate pattern and some spots of arbitrary pattern

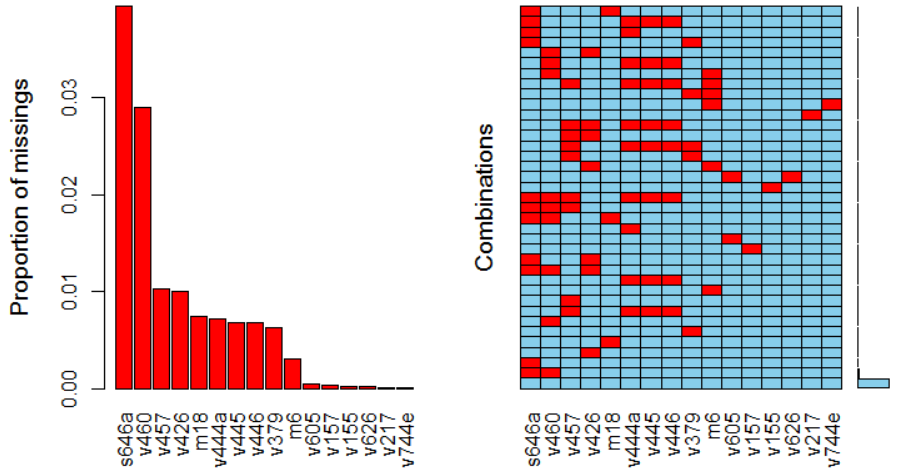


Figure 6. Shows the presence of arbitrary and som spots of univariate pattern

We concluded that there was a presence of both arbitrary and univariate missing data patterns in the under-five child survival data set.

5.3 Missing data mechanisms

Missing data mechanisms bring out the underlying associations between the observed and missing parts of the data. The choice of the missing data approach is also dependent on the unobserved data mechanism. Based on (Rubin, 1987) rules, there are three forms of missing data mechanisms namely; missing at random (MAR), missing completely at random (MCAR) and missing not at random (MNAR). Let Y_i represent incomplete variable, R_i be the response indicator (R_i=1 if Y_i is missing), X_i be the fully observed variable.

5.3.1 Missing at Random

The variable Y_i is said to be MAR if the likelihood of R_i is conditionally dependent on the observed values of Y_i given X_i.

$$P(Y_i/X_i, R_i = 1) = P(Y_i/X_i, R_i = 0) \quad \textbf{(23)}$$

For-example, if we assume that the likelihood of reporting information on immunization status depends on whether the child is alive or dead, then the missing data mechanism for immunization is MAR. Most of the statistical imputation softwares assume that is MAR.

5.3.2 Missing Completely at Random

The variable Y_i is said to be MCAR if the probability of R_i is independent of the unobserved values of Y_i and the values of fully observed variable X_i.

$$P(R_i/Y_i, X_i) = P(R_i) \quad \textbf{(24)}$$

This in simple terms implies that the probability of missing under MCAR doesn't depend on either observed or unobserved values. Statistically, parameter estimates derived with complete case analysis when data is MCAR are unbiased. However the statistical power of the tests will reduce due to reduced sample size. MCAR also leads to larger standard errors (Rubin, 1987). (Allison, 2001) suggests that both MACR and MAR are ignorable missing data mechanisms.

5.3.3 Missing Not at Random

The variable Y_i is said to be MNAR if the distribution of R_i is dependent on the unobserved values of Y_i given X_i i.e the probability of missing depends only on the unobserved data.

$$P(Y_i/X_i, R_i = 1) \neq P(Y_i/X_i, R_i = 0) \quad \textbf{(25)}$$

MNAR is also known as "non-ignorable" missing data. In such a case, (Rubin, 1987) suggested that Y_i and R_i must be modeled jointly together under MNAR assumption for valid inferences to be made.

5.3.4 Checking for missing data mechanism

In this thesis, we checked for the presence of MAR and MCAR assumption. To check for MAR, we created an R indicator variable with 1 if missing and 0 if observed for every variable with missing data. We conducted the two sample t-test for numeric variables (Dong and Peng, 2013) and chi-square test for factor variables.

The two sample t-test statistic tests the null hypothesis

$$H_0 : \mu_1 = \mu_2$$

. It written as

$$t = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad \textbf{(26)}$$

Where s_1^2 and s_2^2 are the corresponding variances.

The chi-square test assesses the association between two groups. It tests the null hypothesis

$$H_0 : \text{There is no association between two groups}$$

It is computed as

$$\chi^2 = \frac{\sum(O_i - E_i)^2}{E_i} \quad \textbf{(27)}$$

Where $E_i = \frac{row\ sums \times column\ sums}{n}$ and the degrees of freedom is $(r-1)(c-1)$.

Table 5 shows results from the MAR tests. The results showed that missingness in 11 covariates was related to the time to event and missingness in four covariates was related to the event variable. The results also shown that MAR was present in 60% of the covariates.

Table 5. MAR assumption

Missing covariate	Prop. missing	Proportion of covariates predicting missingness in the missing covariate data (MAR)		Is MAR related to the response variables (Y=Yes, N=No)	
		MAR Present	MAR Abscent	time to event	event/status
v113	0.05	43.1	56.9	N	N
v221	0.05	66.7	33.3	Y	N
v155	<0.01	7.8	92.2	N	N
v3a00y	0.36	80.4	19.6	Y	Y
v3a00z	0.36	80.4	19.6	Y	Y
v3a08j	0.64	80.4	19.6	Y	N
v426	0.01	47.1	52.9	N	Y
v457	0.01	19.6	80.4	N	N
v503	0.05	66.7	33.3	Y	N
v511	0.05	66.7	33.3	Y	N
v603	0.31	74.5	25.5	Y	N
v604	0.31	74.5	25.5	Y	N
v605	<0.01	7.8	92.2	N	N
v616	0.31	76.5	23.5	Y	N
v741	0.16	70.6	29.4	Y	N
b12	0.7	76.5	23.5	Y	Y
m6	<0.01	17.6	82.4	N	N
m18	0.01	29.4	70.6	N	N
s313b	0.92	66.7	33.3	N	N
v446	0.01	23.5	76.5	N	N
v626	<0.01	3.9	96.1	N	N

We also checked the data for missing completely at random (MCAR) assumption using LittleMCAR test under the BaylorEdPsych package (Beaujean, 2012). The LittleMCAR test assesses the null hypothesis

$$H_0 : Data \quad is \quad MCAR$$

The results from the MCAR tests shown a p value = 0.000 suggesting that data was not MCAR. In conclusion, we assumed that MAR was a plausible assumption for our data.

In this thesis, we apply two multiple imputation approaches namely; Multiple Imputation by Chained Equations (MICE)Buuren and Groothuis-Oudshoorn (2011) and Random Forests (Breiman, 2001) imputation implemented under the MICE framework.

5.4 Multiple Imputation by Chained Equations

Multiple Imputation by Chained Equations (MICE) (Buuren and Groothuis-Oudshoorn, 2011) is a MI approach that imputes data by specifying the imputation model of each variable with incomplete cases by using a set of conditional densities. MICE comes with three good features; 1) accounting for the process that generated the data, 2) mantaining the relations in the data and 3) mantaining the uncertainty about these relations. Because MICE approach specifies an imputation function for each variable with unobserved data, it's appropriate for univariate missing data pattern that requires specific variable distributions i.e predictive mean matching for numeric variables and logistic regression for factor variables (Rubin, 1987; Schenker and Taylor, 1996). MICE further draws imputations based on the Bayesian Gibb's sampling algorithm Geman and Geman (1987) that's appropriate for arbitrary missing data pattern (Little and Rubin, 2014). This then makes MICE approach appropriate for the two types of missing data patterns. The MICE imputation process is described below;

Let $X = (X_1, X_2, ..., X_m)$ be a set partially observed random variables. MICE draws imputations from the unconditional distribution function of X i.e $f(X)$. MICE assumes that the distribution of X is fully specified by a vector of unknown parameters (θ). The first step is to obtain the multivariate distribution of θ. MICE attains the posterior distribution of θ by sampling frequentatively from the conditional densities of the form;

$$\begin{aligned} &P(X_1|X_2, X_3, ..., X_m, \theta) \\ &\qquad\qquad\vdots \\ &P(X_m|X_1, X_2, ..., X_{m-1}, \theta) \end{aligned} \tag{28}$$

$\theta_1, ..., \theta_m$ represent parameters of the respective conditional distributions. Starting with a rough draw from X_{obs}, the t^{th} iteration of the chained equations is a Gibbs sampler (Geman and Geman, 1987; Little and Rubin, 2014) that successively draws

$$\begin{aligned} \theta_1^{*(t)} &\sim P(\theta_1|X_1^{obs}, X_2^{(t-1)}, ..., X_m^{(t-1)}) \\ X_1^{*(t)} &\sim P(X_1|X_1^{obs}, X_2^{(t-1)}, ..., X_m^{(t-1)}, \theta_1^{*(t)}) \\ &\vdots \\ \theta_m^{*(t)} &\sim P(\theta_m|X_m^{obs}, X_1^{(t)}, ..., X_{m-1}^{(t)}) \\ X_m^{*(t)} &\sim P(X_m|X_m^{obs}, X_1^{(t)}, ..., X_{m-1}^{(t)}, \theta_m^{*(t)}) \end{aligned} \tag{29}$$

$X_j^{(t)} = (X_j^{obs}, X_j^{*(t)})$ is the j^{th} imputed random variable at iteration t. Prior imputations/runs $X_j^{*(t-1)}$ enter $X_j^{*(t)}$ through its iteration with other variables. This process is repeated t times to generate t imputations . To achieve m imputations, the process is repeated m times. Below is a description of the MICE procedure in lay terms as described by (Azur et al., 2011) below;

1. Start with a simple a draw say mean as the imputed value for all unobserved variables.
2. Set the simple draw for the individual variable (X) to be imputed to missing
3. Regress the observed values of X on the rest of the variables in the imputation model. X is treated as the outcome and other variables as explanatory.
4. Replace the unobserved values of X with the predictions from the regression model.
5. Repeat steps 2–4 for every incompletely observed variable.
6. To achieve i iteration times, repeat steps 2-5 for every iteration.
7. To achieve m imputations, repeat steps 2-6 to create m multiple imputed data sets.

MICE approach also has the potential to adjust for clustering in the data set by treating the cluster variables as class variables in the imputation function such that imputations are done with in classes. In this thesis, we adjust for the community level (enumeration area) and household level clusters in the imputation model as described in Table 6.

5.4.1 Univariate imputation models used in MICE

Out of the total 23 covariates that contained missing data, 4 were binary, 13 were unordered ($>$ 2 levels), 1 was ordered and 5 were continuous. The idea of chained equations is hinged on the fact that for every variable type, an appropriate univariate or multivariate distribution is used. (Buuren and Groothuis-Oudshoorn, 2011) provides the details of the different distribution that are used in the MICE package that has been applied for this study. They include; logistic regression (logreg) for binary data, multinomial logistic regression (polyreg) for unordered variables with $>$ 2 levels, ordinal/proportional odds logistic regression (polr) for ordered variables with $>$ 2 levels, and predictive mean matching (pmm) for numeric data. Below is a description of the four imputation models;

Logistic regression

Logistic regression (logreg) model combines a set of predictor variables to estimate the likelihood that the event of interest will take place. It estimates the probability of a subject being a member of the category of interest. The predictor can take on any form but the response is a binary category. The logistic mathematical model is is expressed as follows;

$$Log(\frac{P_i}{1-P_i}) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p \quad \textbf{(30)}$$

where β's are the model coefficients, X_s are the predictors, and P_i is the probability of observing a 1 in the response variable. The parameters βs are estimated by using maximum likelihood estimation of the likelihood function constructed from the binomial distribution. An observation is predicted to belong to belong to the class (0,1) with the highest probability.

Multinomial logistic regression

Multinomial logistic regression model is an addition of the binary logistic regression model when the response variable has > 2 nominal levels. The response variable is dummy variable expressed into m $1/0$ indicator variables. Therefore, if there are m nominal levels, there will be $m-1$ indicator variables. The multinomial logistic regression estimates discrete binary logistic regression models for ever $m-1$ dummy variables. Each of $m-1$ model has its own intercept and regression coefficients. The $m-1$ binary logistic models are fitted simultaneously as follows;

$$\begin{aligned} Log\big(\frac{p(y=2)}{p(y=1)}\big) &= \beta_{02} + \beta_{12}X_1 + \beta_{22}X_2 + ... + \beta_{p2}X_p \\ Log\big(\frac{p(y=3)}{p(y=1)}\big) &= \beta_{03} + \beta_{13}X_1 + \beta_{23}X_2 + ... + \beta_{p3}X_p \\ &\vdots \\ Log\big(\frac{p(y=m)}{p(y=1)}\big) &= \beta_{0m} + \beta_{1m}X_1 + \beta_{2m}X_2 + ... + \beta_{pm}X_p \end{aligned} \tag{31}$$

The first category is assumed to be the reference category for each binary model. The parameters $\beta' s$ are estimated by using maximum likelihood estimation of the likelihood function constructed from the multinomial distribution. There are basically m equations that are used to computed the probability that an observation is a member of any of the m categories. An observation is predicted to belong to the category with the highest probability.

Ordinal logistic regression

Ordinal logistic regression is used when the categorical variable has > 2 ordered levels i.e socio-economic status, education level etc. The model is of the form

$$log\big(\frac{p(y \leq c_j)}{p(y > c_j)}\big) = \beta_{0j} + \beta_{1j}X_1 + \beta_{2j}X_2 + ... + \beta_{pj}X_p \tag{32}$$

If the predictors don't depend on the categories but the intercept does, then we have Proportional odds logistic regression

$$log\Big(\frac{p(y \leq c_j)}{p(y > c_j)}\Big) = \beta_{0j} + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p \quad \textbf{(33)}$$

An observation is predicted to belong to the category with the highest probability.

Predictive mean matching (pmm)

Pmm is an imputation approach that uses the usual linear regression model to make predictions (Vink et al., 2014). Under the pmm, the imputations are based on the observed values and it has the potential to preserve the non-linearity even when the structured part of the model used for imputation is not correct. The following is the pmm algorithm used for imputing the missing values.

1. Regress Y_{obs} on X_{obs} to estimate the parameters $\hat{\beta}$,$\hat{\sigma}$, and $\hat{\Sigma}$ by applying the ordinary least squares (OLS) approach. The normal regression model of the form $Y_i = X_i^t\beta + \varepsilon_i$ is applied.
2. Estimate σ_*^2 as $\sigma_*^2 = \hat{\Sigma^T}\hat{\Sigma}/Q$, where Q is a χ^2 with $n-r$ degrees of freedom.
3. Draw β^* from a multivariate normal function centered at $\hat{\beta}$ with variance matrix $\sigma_*^2(X_{obs}^t X_{obs})^{-1}$
4. Calculate $\hat{Y}_{obs} = X_{obs}\hat{\beta}$ and $\hat{Y}_{mis} = X_{mis}\beta^*$
5. For each $\hat{Y}_{mis,i}$, find $D = |\hat{Y}_{obs} - \hat{Y}_{mis,i}|$.
6. Draw one value randomly from $(D^{(1)}, D^{(2)}, D^{(3)})$, where $D^{(1)}, D^{(2)}$ and $D^{(3)}$ are the 3 least objects in the set D and use the matching entry in $Y_{obs,i}$ as the imputed value

5.5 Random forests imputation within the MICE framework

Random forests (Breiman et al., 1984; Breiman, 2001) is a recursive partitioning approach that is applied to predictions and variable selection by splitting the data into homogeneous classes. Recursive partitioning identifies the best predictive split of the outcome variable by searching via all the explanatory variables and picking the variables that provide more gainful information (Breiman, 2001). Random forests imputation algorithm within the mice framework was proposed by (Doove et al., 2014) using the random forests procedure by (Breiman, 2001). The algorithm is described as follows;

Let X be a data matrix. Let X^{obs} and X^{mis} represent the observed and unobserved data. Let p represent the number of variables.

1. For each variable in X_{mis}, start with a simple random draw from the X_{obs}.
2. For each variable in X_{mis}, the first imputation is drawn using the following procedure;
   (a) Draw b training samples using only X_{obs}
   (b) Build one tree on every training sample drawn in step 2(a) using random input selection for every split. For every split, m candidate variables are tried, usually $m = \sqrt{p}$. Gini impurity criterion is used for node splitting. This leads to b trees, where each tree contains many nodes. Each node includes a subset of X_{obs} called donors.
   (c) For cases in X_{mis}, choose which node they will end up according to the b trees resulting into k nodes with donors per member of X_{mis}
   (d) For cases in X_{mis}, take all donors from the k nodes in step 2(c) together and randomly select one X_{obs} value from the donors. Replace X_{mis} values with the imputed values.
3. Repeat step 2 to achieve t iteration times
4. Redo steps 1-3 m times to generate m imputed data sets

This algorithm is inbuilt in the MICE package and is applied to generate multiple imputed data under the random forests imputation strategy. Random forests uses the Gini impurity as a measure of how a randomly selected item from a set would be mis-labeled if it was labeled randomly according to the distribution of labels in the subset. Consider a set of items with R classes, let $i \in 1, .., R$, let f_i be the proportion of an item with label i in the set R. The Gini impurity is computed as

$$GI = \sum_{i=1}^{R} f_i \sum_{k \neq i} f_k = \sum_{i=1}^{R} f_i(1 - f_i)$$

Where $1 - f_i$ is the probability of mistakenly classifying i. The Gini index will be zero if all the items are classified in one class. A split that minimizes the impurity is selected.

Auxiliary variables In addition to the 50 highly ranked covariates by RSF, we included in the imputation model; the demographic characteristics that were not highly ranked, survey sample weighting variable and the two cluster variables.

5.6 Set up of the imputation model

We set up the imputation model using the *quickpred()* function introduced by (Buuren and Groothuis-Oudshoorn, 2011). The *quickpred()* function allows specification of the variables to be included for every variable with missing data, minimum correlation to be used, and any variables to be excluded. To improve the quality of our imputations, we set the minimum correlation as 10% and this allowed about 10 to 15 covariates to be included in the imputation model for every covariate. To further improve the imputations, we included all the auxiliary variables, time to event and event variables as predictors in the predictor matrix. Table 6 describes the different imputation strategies used.

Table 6. Description of the imputation strategies (HH-Household, EA-Enumeration Area, RF-random forest)

Imputation Strategy	Description
MICE Flat	We included in the predictor matrix all the covariates that had at least 10% correlation with the missing values in a missing covariate and observed values in another variable. The event, time, and auxiliary variables were treated like other covariates in the imputation model. For every variable type, we used the mice inbuilt models described in this section
MICE HH	We included in the predictor matrix all the covariates that had at least 10% correlation with the missing values in a missing covariate and observed values in another variable. The event, time, and auxiliary variables were treated like other covariates in the imputation model. For every variable type, we used the mice inbuilt models described in this section. We treated the household cluster variable as a class variable in the imputation model.
MICE EA	We included in the predictor matrix all the covariates that had at least 10% correlation with the missing values in a missing covariate and observed values in another variable. The event, time, and auxiliary variables were treated like other covariates in the imputation model. For every variable type, we used the mice inbuilt models described in this section. We treated the enumeration areas cluster variable as a class variable in the imputation model.
MICE RF	We included in the predictor matrix all the covariates that had at least 10% correlation with the missing values in a variable and observed values in another variable. The event, time, and auxiliary variables were treated like other covariates in the imputation model. The method for imputation was specified as random forests ("rf") under the mice package. We used ntree=100.

5.7 Creation of multiple imputed data sets

Deciding on the number of imputations is dependent on the proportion of missing information in the data set. (Rubin, 1987) suggested a formula that is based on relatively efficiency (RE) that can be used to determine the number of imputations required to achieve a relatively higher efficiency.

$$RE = [1 + \frac{\gamma}{m}]^{-1}$$

Where γ is the proportion of information missing due to the missing data. Prior reviews suggest that imputations ranging from 5 to 10 are adequate to generate plausible imputations (Schafer and Olsen, 1998). We considered 10 imputations and 20 iterations per imputation strategy as appropriate for our thesis leading to 10 multiple imputed data sets per imputation strategy.

5.8 Selection of the best imputation strategy

To identify the imputation strategy to provided better imputations, we conducted diagnostic checks on the quality of imputations, analyzed imputed data using random survival forests and Cox regression model. We examined the performance of the imputation strategies based on the average estimated parameters, probability values, confidence intervals, and the corresponding standard errors.

5.8.1 Assessing convergence

To examine the convergence of the Gibb's sampling algorithm, we examined the parallel imputation streams for the mean and standard deviation of the imputation using the convergence plots introduced by Buuren and Groothuis-Oudshoorn (2011). For healthy convergence to be realized, the resulting streams must freely intermix with each other with no explicit directions, that is to say that the variations between cycles is no greater than the variations with each individual cycle.

5.8.2 Diagnostic checking

Imputed data should ideally exhibit the same distributional properties as the observed data values. We used summary statistics and frequency distributions of both imputed and original data for some variables to assess the closeness of the imputed data to the original data. We used density plots of different variables as a diagnostic measure of the quality of the imputation strategies. Density plot show marginal distributions of the observed vs. imputed values for a given variable. Large differences may imply that the observed values could not provide enough information to provide plausible imputations or the imputation model used was not appropriate.

5.8.3 Statistical inference of the imputed data

After conducting diagnostic checks on the imputed data, we conducted separate statistical inference analysis on the imputed data for each imputation strategy. We analyzed the imputed data using random survival forests (RSF) and Cox regression model. RSF model procedures are already described in our previous sections. The Cox regression analysis was based on the Rubin rules (Rubin, 1987) that involve two steps; 1) analyzing individual (m) data sets separately and 2) the pooling results from the individual data sets together.

Rubin analysis procedures

Let H denote the coefficient estimate from the Cox regression model. Let $\hat{H}_l$ denote the estimate from the l^{th} imputed data. Let $\hat{F}_l$ denote the estimate of the variance of $\hat{H}_l$ for l^{th} imputed data.

The pooled estimate of H is given by

$$\bar{H} = \frac{1}{m}\sum_{l=1}^{m}\hat{H}_l \tag{34}$$

$\hat{H}_l$ accounts for sampling uncertainty. $\bar{H}$ accounts for both the sampling and missing data uncertainty.

The variance of $\bar{H}$ is composed of ; the with-in-imputation variance ($\bar{F}$) and the between-imputation-variance (G).
The with-in-imputation variance ($\bar{F}$) is computed as follows

$$\bar{F} = \frac{1}{m}\sum_{l=1}^{m}\bar{F}_l \tag{35}$$

$\bar{F}_l$ is the l^{th} imputation variance of $\hat{H}_l$.

The variance between the m-imputed data sets (G) is computed as follows

$$G = \frac{1}{m-1}\sum_{l=1}^{m}(\hat{H}_l - \bar{H})(\hat{H}_l - \bar{H})^t \tag{36}$$

The resulting total variance is estimated as follows

$$\begin{aligned} T = &\quad \bar{F} + G + \frac{G}{m} \\ = &\quad \bar{F} + (1 + \frac{1}{m})G \end{aligned} \tag{37}$$

The term $\frac{G}{m}$ is the imputation error, the extra imputation variance arising from the fact that $\bar{H}$ is based on finite m.

Variance ratios

1. Proportion of variance attributable to the missing data

$$\lambda = \frac{G + G/m}{T} \tag{38}$$

2. Relative increase in variance due to non-response

$$r = \frac{G + G/m}{\bar{F}} \tag{39}$$

Relationship between r and λ

$$r = \frac{\lambda}{1-\lambda} \tag{40}$$

3. Fraction of information about H missing due to missing data

$$\gamma = \frac{r + 2/(v+3)}{1+r} \tag{41}$$

v is the degrees of freedom

Relationship between γ and λ

$$\gamma = \frac{v+1}{v+3}\lambda + \frac{2}{v+3} \tag{42}$$

Statistical inference for $\bar{H}$

1. Confidence Interval

The $100(1-\alpha)\%$ confidence interval of $\bar{H}$ is computed as

$$\bar{H} \pm t(v, 1-\alpha/2)\sqrt{T} \tag{43}$$

Where $t(v, 1-\alpha/2)$ is the t-distribution quantile with v degrees of freedom and α level of significance.

2. P value

From the test of the null hypothesis that $H{=}H_0$, we estimate the p-value as follows

$$P_j = Pr\Big[F_{1,v} > \frac{(H_0 - \bar{H})^2}{T}\Big] \tag{44}$$

H_0 is the specified value of H, $\bar{H}$ is the pooled estimate, v is the degrees of freedom, T is total variance

Computation of the degrees of freedom

The degrees of freedom V_m or the adjusted V_m^* is computed by the following formula.

$$\begin{aligned} V_m &= (m-1)[1 + \frac{1}{r}]^2 \\ &= \frac{m-1}{\lambda^2} \end{aligned} \tag{45}$$

V_m assumes $n = \infty$
The adjusted formula V_m^*

$$V_m^* = \frac{V_m * V_0^*}{V_m + V_0^*} \tag{46}$$

V_0^* is the estimated observed data degrees of freedom that accounts for missing data. V_0^* is computed as

$$V_0^* = \frac{V_0 + 1}{V_0 + 3} \tag{47}$$

Where $V_0 = n - k$

Cox-Proportional Hazards model

We applied the Cox-Proportional Hazards regression model (Cox, 1972) to study the effects of covariates on the time-to-event response variable. Th formula for the Cox model of an observation given covariates $(X_1, ..., X_m)$ is as follows

$$\begin{aligned} \lambda(t|X) &= \lambda_0(t) \exp(\beta_1 X_1 + ... + \beta_m X_m) \\ &= \lambda_0(t) \exp(\beta^t X) \end{aligned} \tag{48}$$

Where $\lambda_0(t)$ is the baseline hazard, and assumes no distribution and doesn't depend on the values of covariates. Consider two observations with covariates X and X^*, the ratio of their hazards is

$$\begin{aligned} HR(t) &= \frac{\lambda(t|X)}{\lambda(t|X^*)} \\ &= \frac{\lambda_0(t) \exp(\beta^t X)}{\lambda_0(t) \exp(\beta^t X^*)} \\ &= \exp\Big(\sum_{i=1}^{m} \beta_i (X - X^*) \Big) \end{aligned} \tag{49}$$

$HR(t)$ is known as the hazard ratio and compares the hazard of having an event with covariate value X to the hazard of having an event with covariate value $X*$. The proportional hazards assumption assumes that $HR(t)$ is constant over time.

To estimate the parameter estimates (β), we obtain the maximum likelihood estimates of the Cox-partial likelihood function. We use a partial likelihood function as opposed to the usual likelihood function since the function is based on the event set only.
Let $T_1 < T_2 < ... < T_D$ be distinct ordered event times, Let i represent the individual with

an event $T(i)$, let $R(t)$ denote the risk set at time T.
The Cox-partial likelihood is given by

$$L(\beta) = \prod_{j=1}^{D} \Big[\frac{\exp(\beta^t X_j)}{\sum_{i \in R(T)} \exp(\beta^t X_i)} \Big] \tag{50}$$

The log-partial likelihood

$$log\Big(L(\beta)\Big) = \sum_{j=1}^{D} \Big[\beta^t X_j - log\Big(\sum_{i \in R(T)} \exp(\beta^t X_i) \Big) \Big] \tag{51}$$

Maximizing the $log\Big(L(\beta)\Big)$ by solving equation (52) gives β

$$\frac{\partial log\Big(L(\beta)\Big)}{\partial \beta_i} = 0 \tag{52}$$

The variance of β is computed as $var(\beta) = I^{-1}$ where I is the Fisher-information matrix computed as follows $I = E\Big[\Big(\frac{\partial log L(\beta)}{\partial \beta_i}\Big) \times \Big(\frac{\partial log L(\beta)}{\partial \beta_j}\Big)\Big]$

From the β, we derive the Wald and Likelihood ratio tests of global hypothesis

$$H_0 : \beta = \beta_0$$

Under H_0, we get
1. Wald test

$$\chi^2_w = (\beta - \beta_0)^t I(\beta)(\beta - \beta_0) \ \chi^2_p \tag{53}$$

2. Partial Likelihood ratio

$$\chi^2_{LR} = 2(log(L(\beta)) - 2log(L(\beta_0)) \ \chi^2_p \tag{54}$$

The 95% CI for β is computed as

$$\Big[\beta - Z_{\alpha/2} se(\beta), \beta + Z_{\alpha/2} se(\beta)\Big] \tag{55}$$

Where $alpha$ is the significance level and Z is the statistic from the normal distribution.
The 95% CI for hazard ratio is constructed as

$$\exp\Big[\beta - Z_{\alpha/2} se(\beta), \beta + Z_{\alpha/2} se(\beta)\Big] \tag{56}$$

In case of ties, we apply the Breslow approximation of the $L(\beta)$ as follows;

$$L(\beta) = \prod_{j=1}^{D} \Big[\frac{\exp(\beta^t) \sum_{k \in D_j} X_j)}{\Big(\sum_{i \in R(T)} \exp(\beta^t X_i)\Big)^{d_j}} \Big] \tag{57}$$

Testing the proportional hazards assumption

We assessed the proportional hazards assumption using the statistical test based on the scaled Schoenfeld residual tests under the survival package in R. For every variable, the test correlates the scaled Schoenfeld residuals with time to assess the independence of residuals against time.
From the partial likelihood, the parameter β are estimated from

$$\sum_{i=1}^{d} \Big(x_i - E[x_i | R(t_i)] \Big) \quad = 0 \tag{58}$$

Where

$$E[x_i | R(t_i)] = \frac{\sum_{i \in R(t)} x_i \exp(x_i^t \beta)}{\sum_{i \in R(t)} \exp(x_i^t \beta)}$$

The schoenfeld residuals are defined as follows;

$$r_i = x_i - E[x_i | R(t_i)] \tag{59}$$

The plot of r_i against the ranks of survival times is used to assess violations from the PH assumption. For PH to be met, the line representing the coefficients should be a horizontal line since hazard ratio is constant over time. Systematic deviations from the horizontal line are suggestive of the PH violation.

5.9 Results

5.9.1 Distribution of missingness of covariates by demographic characteristics

Table 7 indicates the relationship between missingness of some of the partially observed variables with the demographic characteristics and response variables. The results show that other than age of the respondent, missingness on succeeding birth interval (b12) was significantly related with observed values of the rest of the demographic characteristics and response variables. Data on succeeding birth interval was more likely to be observed for observations with long time to event data than those with short time to event (42 vs. 23). The information on succeeding birth interval was also more likely to be missing for the alive children than the dead children (79% vs. 41%). Missing data on succeeding birth interval was also more likely to happen for female headed households than the male headed households (77% vs. 68%), for those residing in urban areas than rural areas (79% vs. 67%), for the rich than the poor (77% vs. 64%) and for the single than the married (92% vs. 68%). Additionally, missing data on succeeding birth interval was more likely to happen for the observations with secondary level than those with no education (77% vs.

63%) and for those employed than the unemployed (75% vs. 71%). Missingness in source of water was also significantly related to to time, respondent's education, age, occupation, wealth index, and marital status. Similarly, missingness in preferred waiting time for a child was also significantly related to time,sex of household head, respondent's education, age, occupation, wealth index, and marital status.

Table 7. Relationship between missingness in covariates and demographic characteristics

Characteristics	Partially observed covariates [Obs. = Observed | Mis. = Missing | *Significant at 5% level of significance]								
	Succeeding birth interval (b12)			Source of drinking water (v113)			Preferred waiting time for a/ another child		
	Obs.	Mis.	P value	Obs.	Mis.	P value	Obs.	Mis.	P value
Response variables									
Time (mean)	42.3	22.6	0.000*	28.6	26.1	0.003*	27.6	30.5	0.000*
Event status (%)									
Alive	29.1	70.8	0.000*	95.4	4.6	0.276	69.1	30.9	0.658
Dead	58.6	41.4		93.9	6.1		70.3	29.7	
Demographic variables									
Place of residence(%)									
Urban	21.3	78.7	0.000*	94.9	5.1	0.306	67.7	32.3	0.112
Rural	32.5	67.5		95.4	4.6		69.5	30.5	
Sex of Household head(%)									
Male	31.4	68.6	0.000*	94.9	5.1	0.306	85.7	78.8	0.000*
Female	22.7	77.3		95.4	4.6		14.3	21.2	
Respondent's education level(%)									
None	36.7	63.3	0.000*	96.1	3.9	0.000*	66.9	33.1	0.000*
Primary	29.5	70.5		95.8	4.2		66.3	33.7	
Secondary	22.8	77.2		92.8	7.2		81.2	18.8	
Higher	27.1	72.9		90.6	9.4		76.5	23.5	
Respondent's age	29.4	29.4	0.863	29.5	26.8	0.000*	27.1	34.5 (mean)	0.000*
Respondent's occupation (%)									
No work	29.3	70.7	0.000*	93.6	6.4	0.000*	74.8	25.2	0.000*
Employee	25.2	74.8		93.6	6.4		72.5	27.5	
Self-employed	31.0	69.0		96.1	3.9		66.8	33.9	
Wealth index (%)									
Poorer/Poor	36.0	64.0	0.000*	96.2	3.8	0.001*	68.7	31.3	0.002*
Middle	29.5	70.5		94.9	5.1		66.5	33.5	
Rich/Richer	23.1	76.9		94.5	5.5		71.0	29.0	
Marital status (%)									
Never	8.0	92.0	0.000*	95.5	4.5	0.000*	79.2	20.8	0.000*
Married	32.2	67.8		95.6	4.4		70.2	29.8	
Widow	22.8	77.2		91.8	8.2		30.4	69.6	
Divorced	22.2	77.8		92.9	7.1		60.9	39.1	

5.9.2 Log-rank test of the observed vs. missing

Table 8 shows results from the two sample log-rank test that compared the survival time of the observed vs. missing. The results shows that survival difference was significantly different in only four covariates (19%).

Table 8. Results from the two sample log-rank test (observed vs. missing)

Covariate	P value	Covariate	P value
v113	0.189	v221	0.265
v155	0.818	v3a00y	0.009*
v3a00z	0.009*	v3a08j	0.751
v426	0.007*	v457	0.857
v503	0.277	v511	0.265
v603	0.428	v604	0.478
v616	0.491	v741	0.298
b12	0.000*	m6	0.258
m18	0.921	s316b	0.634
v446	0.365	v626	0.804
v705	0.718		

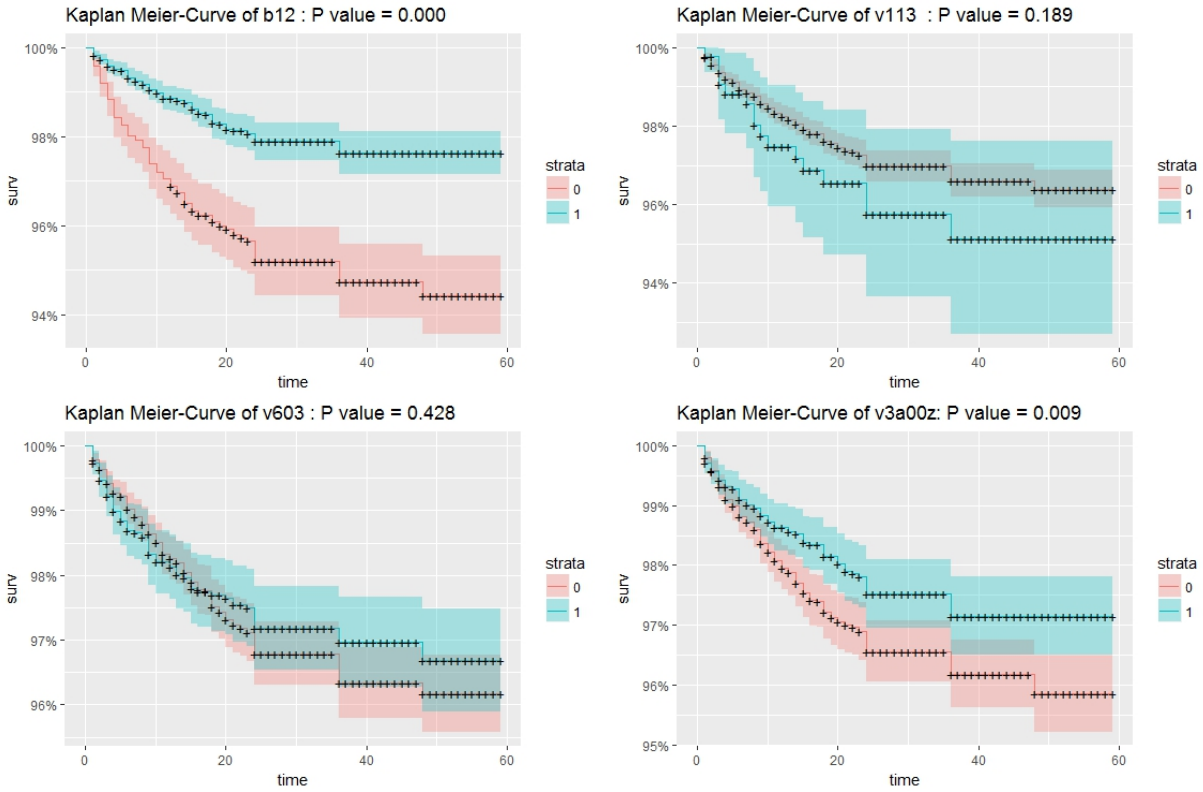


Figure 7. Kaplan Meier plots of covariates stratified by missing and observed. 0=Observed and 1=missing. For-example the KM curves for b12 (suceeding birth interval) shows that survival time was significantly higher for the missing cases than for the observed cases while the KM curve for v113 shows that survival time was higher-not significant for the observed cases than for the missing cases

5.9.3 Implementing multiple imputation

Out of the total 23 covariates with missing data in the imputation model, 3 covariates were not imputed across the four imputation approaches. Two of these covariates had over 60% missing data. One covariate that had 36% missing records was only imputed for MICE RF and MICE EA imputation strategies. The proceeding analysis doesn't include these variables.

5.9.4 Convergence of the imputation iterations

Table 9 shows how the variables converged after the several iterations. For convergence to be achieved, no parallel streams should be observed in the imputations. The results indicate that random forests achieved better convergence (90%) than the rest of the imputation strategies. Treating household numbers and enumeration areas as classes in the imputation model showed no large effect on the convergence.

Table 9. Convergence of the imputation iterations

Imputation strategy	Health Convergence	Unhealthy Convergence
MICE Flat Imputation (%)	55	45
MICE HH Imputation (%)	60	40
MICE EA Imputation (%)	55	45
RF Imputation (%)	90	10

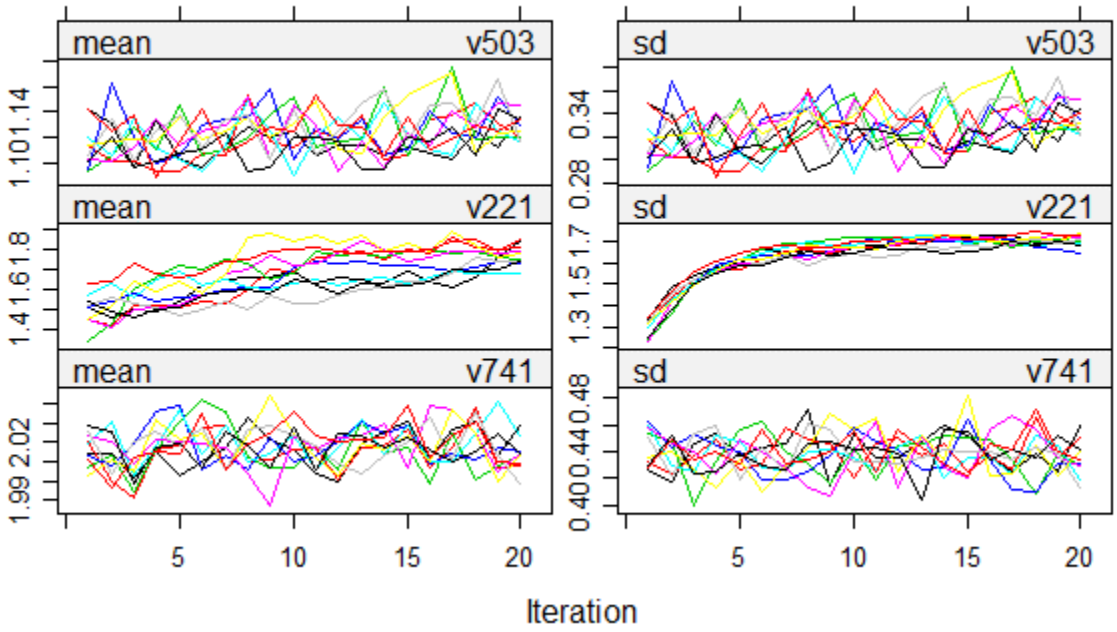


Figure 8. Both scenarios of healthy and unhealthy convergence. Variables v741 and v503 show health convergence while variable v221 shows the unhealthy convergence

5.9.5 Comparison of the marginal distributions of the imputed vs. observed

The study used density plots to check the marginal distributions of the observed vs. imputed. The results indicate that MICE RF imputation strategy produced imputations that were more closely similar with the observed for most of the variables (56%) compared to the other imputation strategies. There are no large observed differences for the flat imputation and the two strategies that treated clusters as classes in the imputation model.

Table 10. Comparison of marginal density plots of imputed vs. observed values by imputation strategy

Imputation strategy	Comparison of marginal density plots of imputed vs. observed		
	Completely identical	Fairly identical	Completely non-identical
MICE Flat Imputation (%)	17	39	44
MICE HH Imputation (%)	22	33	44
MICE EA Imputation (%)	17	39	44
RF Imputation (%)	56	22	22

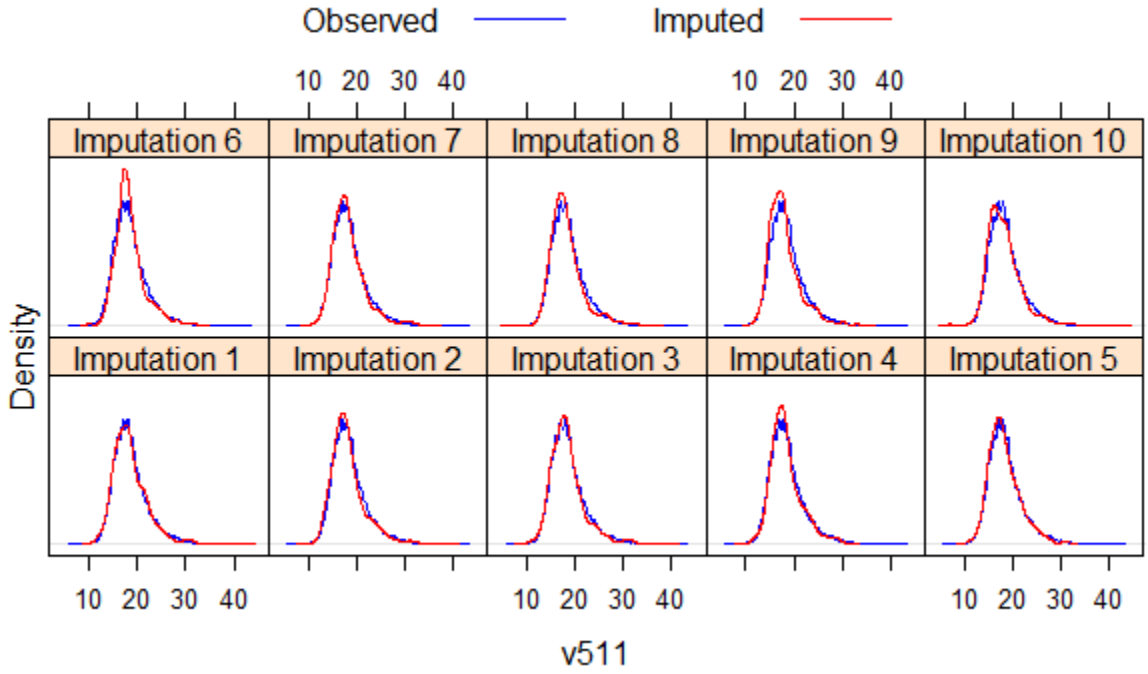


Figure 9. A case of a variable with a completely identical marginal distribution for observed vs. imputed values. Variable v511 achieved completely identical marginal distributions for all the imputation strategies

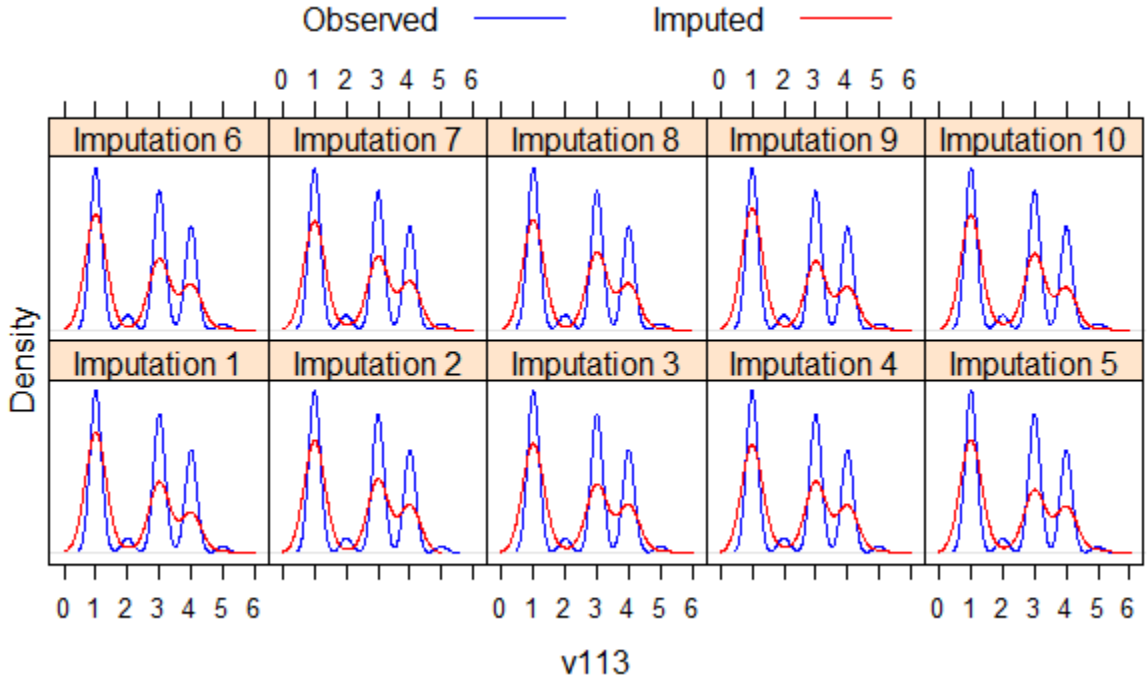


Figure 10. Variable v113 shows a case of a variable with a fairly-completely identical marginal distribution for observed vs. imputed values.

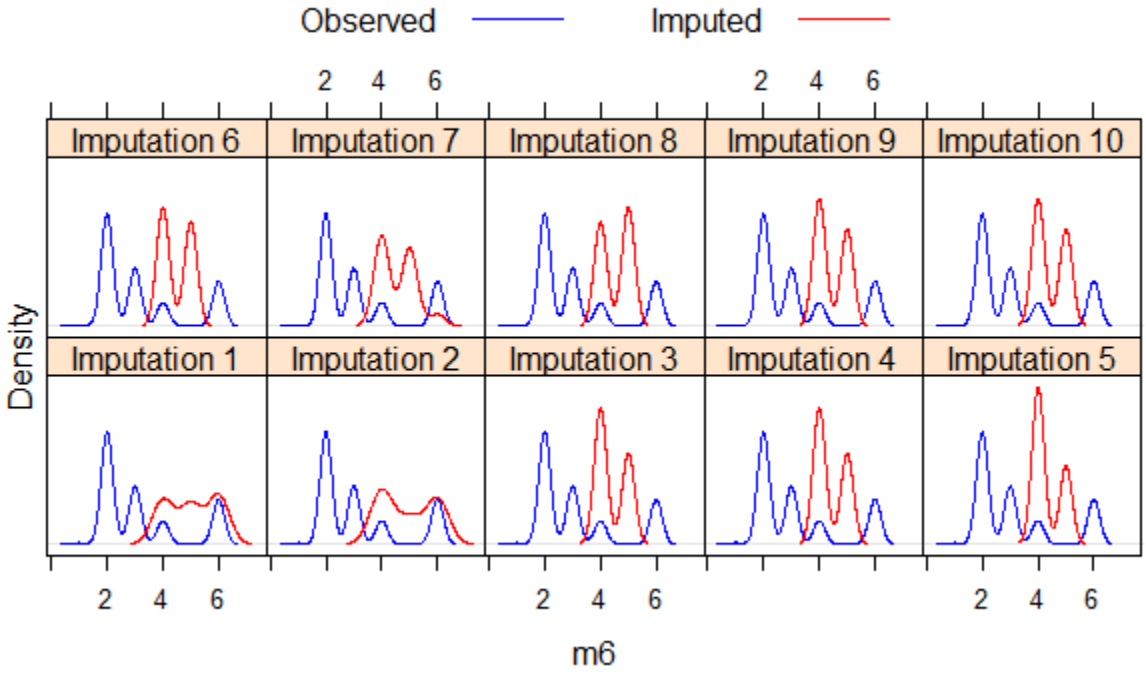


Figure 11. Variable m6 shows a case of a variable with a non-completely identical marginal distribution for observed vs. imputed values.

5.9.6 Summary statistics of the imputed data sets

The study analyzed the data from multiple imputed data sets to compare the means and proportions of the variables with missing data. Table 11 shows that overall the standard errors of the means and proportions were smaller after multiple imputation compared to the complete case analysis. For continuous variables, the results from the imputation strategies were closely related except for one continuous variable (v616) that contained an outlier and RF approach produced mean much closer to the observed mean. The mean (SE) at complete case was 83.2(3.32), 115.1(1.01) for MICE Flat, 120.1(1.03) for MICE HH, 113.2(1.00) for MICE EA and 66.4(0.78) for MICE RF. For categorical variables, the findings from the four imputation strategies were closely related for all the imputation strategies except for variable v3a00z that had a larger proportion of missing data (36%) where random forests produced estimates much closer to the true estimates. The results suggest that with low proportion of missingness, any of the imputation strategies can be used. However, in cases of high level of missingness and outliers, random forests may perform better. Overall, there are no large observed differences from the estimates from the two MICE strategies that treated household and community clusters as class variables and the MICE Flat strategy.

Table 11. Means and proportions with the corresponding standard errors for covariates with missing data by imputation strategy

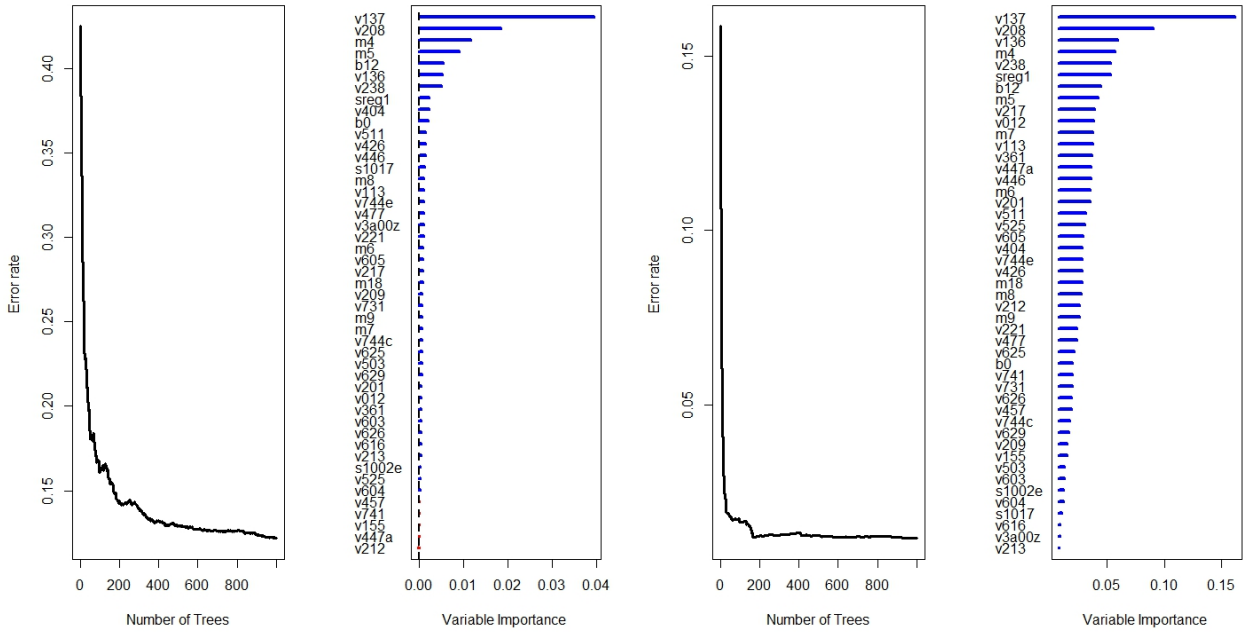
Variable	Missing (%)	Comparison of summaries/proportions per imputation strategy against the complete case (CC) analysis				
		CC	MICE Flat Imputation	MICE HH Imputation	MICE EA Imputation	MICE RF Imputation
Succeding birth interval (b12), mean (SE)	70	28.1(0.16)	25.3(0.03)	23.6(0.03)	25.6(0.03)	25.3(0.02)
v3a00z, % (SE)	36					
No		0.08(0.003)	0.21(0.001)	0.11(0.001)	0.24(0.001)	0.06(0.001)
Yes		0.92(0.003)	0.79(0.001)	0.88(0.001)	0.76(0.001)	0.94(0.001)
v604, % (SE)	31					
<= 12 months		0.10(0.004)	0.14(0.001)	0.16(0.001)	0.14(0.001)	0.08(0.001)
1 years		0.30(0.006)	0.33(0.001)	0.33(0.001)	0.32(0.001)	0.44(0.002)
2 years		0.52(0.006)	0.36(0.001)	0.36(0.001)	0.36(0.001)	0.41(0.002)
3 years		0.06(0.003)	0.04(0.001)	0.04(0.001)	0.04(0.001)	0.04(0.011)
4 years		0.03(0.002)	0.12(0.001)	0.11(0.001)	0.13(0.001)	0.02(0.000)
v616, mean (sd)	31	83.2(3.32)	115.1(1.01)	120.1(1.03)	113.2(1.00)	66.4(0.78)
v503, % (SE)	5					
Once		0.84(0.004)	0.84(0.001)	0.84(0.001)	0.84(0.001)	0.84(0.001)
More than once		0.16(0.004)	0.16(0.001)	0.16(0.001)	0.16(0.011)	0.16(0.001
v511, mean (sd)	5	18.54(0.037)	18.53(0.011)	18.53(0.011)	18.53(0.011)	18.53(0.011)
v113, % (SE)	5					
Piped water		0.38(0.005)	0.38(0.001)	0.38(0.001)	0.38(0.001)	0.39(0.001)
Borehole		0.03(0.002)	0.03(0.001)	0.03(0.001)	0.03(0.001)	0.03(0.001)
Well		0.33(0.005)	0.32(0.001)	0.32(0.001)	0.32(0.001)	0.32(0.001)
Surface water		0.24(0.004)	0.24(0.001)	0.24(0.001)	0.24(0.001)	0.24(0.001)
Others		0.01(0.001)	0.01(0.000)	0.015(0.000)	0.01(0.000)	0.01(0.000)

5.9.7 Application of random survival forests on the imputed data sets

We applied random survival forests model on the imputed data from the different strategies and compared the variable importance scores. RSF Before represents the RSF model fit using the original data set applying the RSF inbuilt measures of treating missing data. The results indicate that the prediction error rate from the imputed data was much smaller (1.00%-1.16%) compared to the error rate from the RSF fit on the observed data (12.34%). The results also indicate that the VarImp scores were similar across the four imputation strategies. However, the VarImp scores from the RSF model on the observed data set were very small. Basing on the VarImp threshold of 0.002 (Ishwaran et al., 2008) that suggests that covariates with VarImp of less than 0.002 are regarded as noisy and thus less-predictive, our results indicate that all the 47 predictors after multiple imputation were highly predictive. Basing on the RSF model fit on the observed data, only 10 predictors were highly predictive. Even though the importance scores are largely different, the ranking of the risk-factors was related with minimal discrepancies as shown in Figure 12. Basing on the ranking of the RSF model for the observed data set, the following risk-factors were highly predictive of under-five child mortality in Tanzania. They include v137-number of under-five children in the household, v208-number of births in the last 5 five years, m4-child's breast feeding history, m5-number of months child breast fed, v238-number of under three children, v136-total number of household members, b12-succeeding birth interval, sreg1-country area zones, v404-currently breast feeding, b0-child's birth status.

Table 12. Random survival forests variable importance scores by imputation strategy.

Variable	VarImp scores by imputation strategy				
	RSF Before	MICE Flat	MICE HH	MICE EA	MICE RF
v137	0.039	0.156	0.16	0.156	0.162
v208	0.018	0.088	0.090	0.087	0.090
m4	0.012	0.055	0.054	0.052	0.057
m5	0.009	0.041	0.040	0.041	0.042
b12	0.005	0.036	0.035	0.040	0.044
v136	0.005	0.059	0.057	0.056	0.059
v238	0.005	0.052	0.052	0.053	0.053
sreg1	0.002	0.052	0.051	0.052	0.053
v404	0.002	0.027	0.029	0.031	0.028
b0	0.002	0.019	0.019	0.019	0.020
v511	0.001	0.031	0.031	0.031	0.031
v426	0.001	0.028	0.028	0.028	0.028
v446	0.001	0.034	0.034	0.033	0.035
s1017	0.001	0.010	0.010	0.010	0.010
m8	0.001	0.027	0.027	0.027	0.027
v113	0.001	0.035	0.035	0.034	0.037
v744e	0.001	0.027	0.026	0.025	0.028
v477	0.001	0.022	0.022	0.022	0.023
v3a00z	0.001	0.005	0.006	0.004	0.009
...					
Error rate	12.34%	1.00%	1.03%	1.02%	1.16%



(a) Variable ranking based on the original data set **(b) Variable ranking based on the imputed data from the random forests approach**

Figure 12. Shows random survival forests under-five risk factor ranking of the 47 risk-factors based on both the imputed data and original data. The risk factor ranking is comparable with minimal discrepancies observed. The prediction error rate from the imputed data set is smaller than the error rate from the original data set. The ranking of the imputed data sets from the three MICE strategies was very closely similar to the ranking from the random forests approach

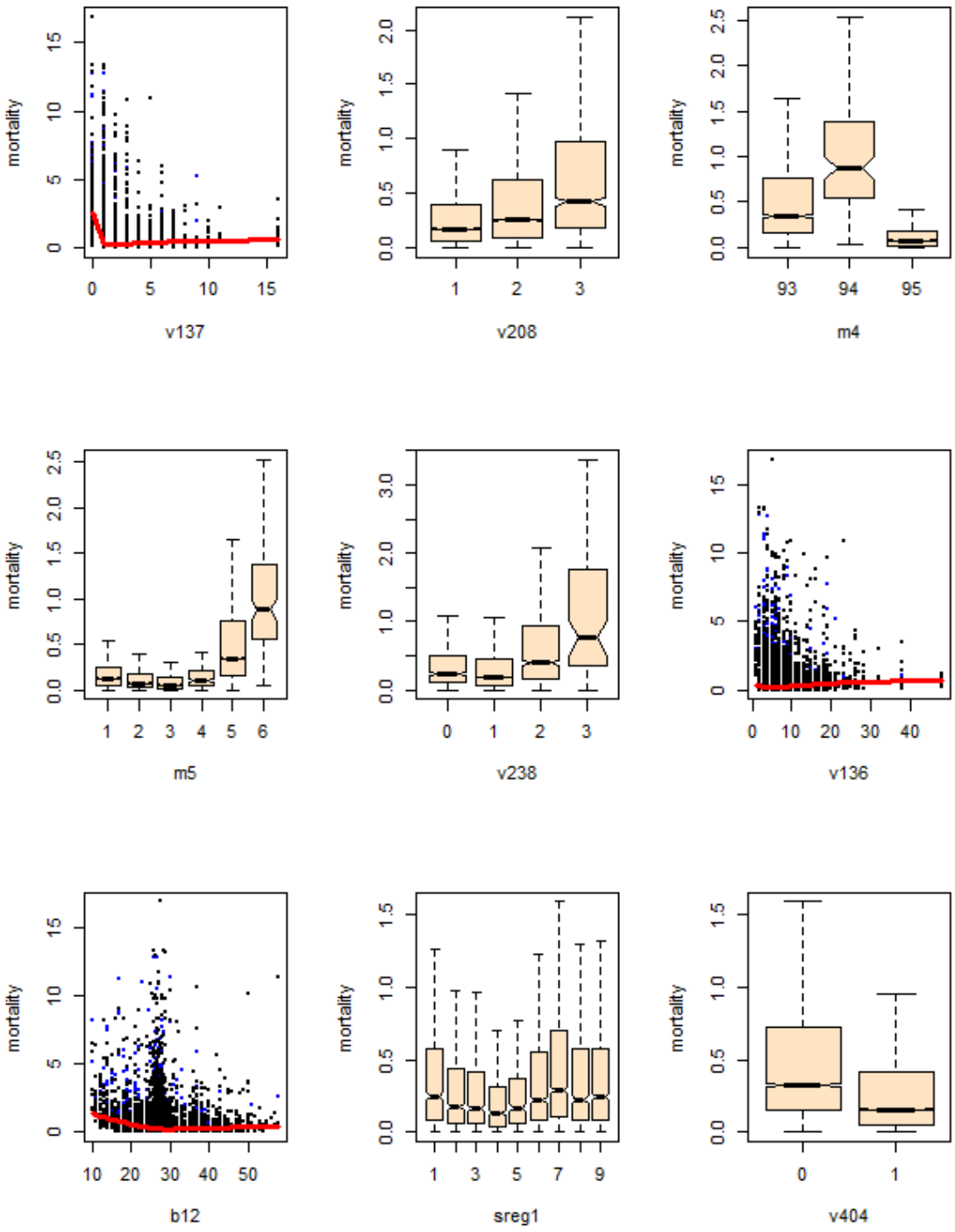


Figure 13. Partial plots for top nine predictors selected using the observed data set while using inbuilt random survival forests missing data handling technique. Values on the vertical axis show expected mortality for predictor, after correcting for all other predictors. Dashed lines are for numeric predictors and box plots are for categorical variables. For example, the plot for b12-succeeding birth interval shows that mortality was decreasing with increasing birth interval. Plot for v208-birth in the last years shows that mortality was increasing with increase in the number of births. The plot for m4 shows that mortality was higher in categories 94(never breast fed), 93 (ever breastfed) and 95 (still breastfeeding) respectively

5.9.8 Testing Proportional Hazards assumption

We first checked for PH assumption on each of the 47 covariates before and after multiple imputation. We plotted the scaled Schoenfeld residuals against time and also conducted the scaled Schoenfeld residual proportional hazards test to assess for PH assumption. Figures 14 and 15 show plots of the proportional hazards tests for two of the covariates. For PH to be met, the Schoenfeld residual test must return a P value greater than 0.05 and the residuals plot must be a model estimates are constant over time (i.e should lie along the horizontal line).

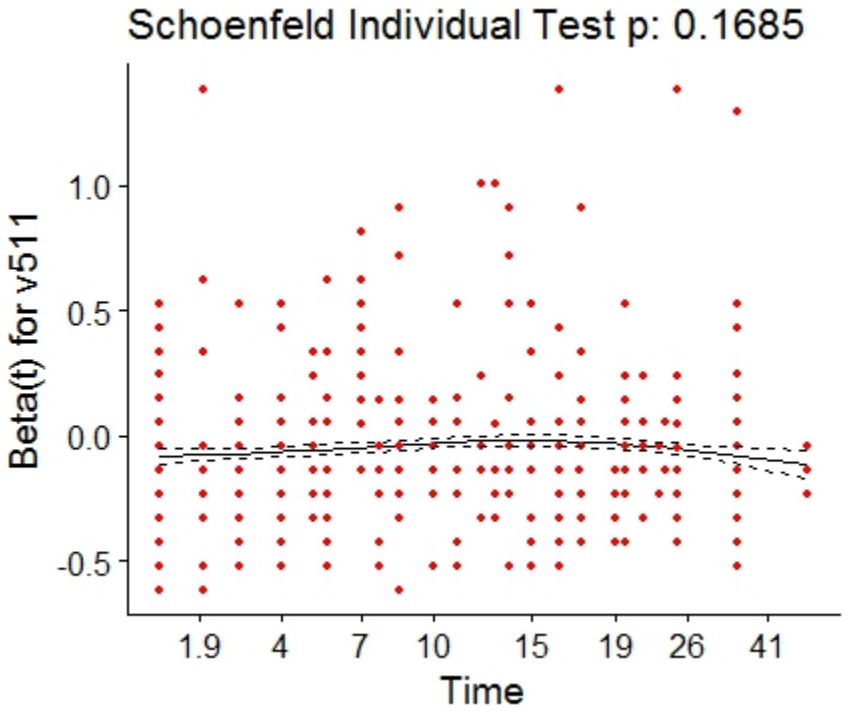


Figure 14. A case of a covariate satisfying PH assumption

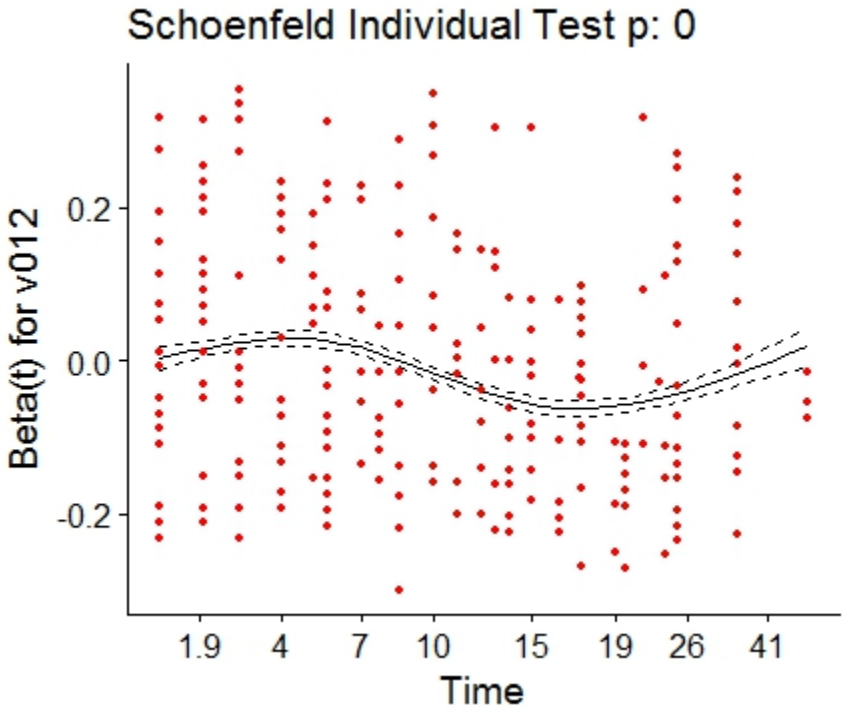


Figure 15. A case of a covariate that doesn't satisfy PH assumption

Table 13 shows a summary of results from the PH assumption. The results indicated that a total of 34 covariates were satisfying the PH assumption before multiple imputation and only 10 covariates were satisfying the PH assumption after imputing data. This result was the same across all the imputation strategies. This result suggests that multiple imputation affected the proportional hazards assumption in the data.

Table 13. Testing for Cox Proportional Hazards assumption before and after multiple imputation

Description	Imputation Strategy				
	CC	Mice Flat	Mice HH	Mice EA	Mice RF
Met PH ,n (%)	34(0.74)	10(0.22)	10(0.22)	10(0.22)	10(0.22)
Failed, n (%)	12(0.26)	36(0.78)	36(0.78)	36(0.78)	36(0.78)

5.9.9 Multivariate Cox-PH regression analysis [Rubin's analysis]

Table 14 shows findings from the multivariate Cox regression analysis. The results indicated that estimates from the complete case analysis were largely different from the estimates from the multiple imputed data sets. The differences in results of the complete case

analysis and multiple imputation are attributable to the significant reduction in sample size under the complete case analysis. A total of 6,525 observations out of 9,779 were deleted from the complete case analysis due to missingness. This result confirms the undesirable effects of complete case analysis in the presence of missing data. The table further reveals that the effect of the covariates v426, v525, and v604 on the response was altered after multiple imputation. The findings from the three MICE strategies (MICE Flat, MICE HH, MICE EA) are comparably similar. There are some observable differences in the estimates of MICE RF and other three MICE approaches. A thorough comparison of the imputation strategies is done based on the Univariate Cox regression model shown in Table 15.

Table 14. Multivariate Cox regression analysis of imputed data from the four imputation strategies

Covariate	Missing(%)	Complete Case	MICE Flat		MICE HH		MICE EA		MICE RF	
		Est,SE	Est,SE	R	Est,SE	R	Est,SE	R	Est,SE	R
v511(β_1)	5%	0.00,0.036	-0.01,0.022	0.021	-0.01,0.022	0.034	-0.01,0.022	0.040	-0.01,0.022	0.034
v626(β_1)	0.02%	NA,0.000	-0.27,0.301	0.527	-0.15,0.210	0.258	-0.31,306	0.506	-0.06,0.170	0.003
v626(β_2)		0.12,0.230	0.20,0.153	0.012	0.21,0.153	0.021	0.22,0.153	0.013	0.14,0.158	0.014
V426(β_1)	1%	-0.39,0.532	-0.02,0.270	0.006	-0.02,0.270	0.021	-0.02,0.269	0.020	-0.01,0.270	0.001
V426(β_2)		0.18,0.223	0.16,0.136	0.010	0.15,0.137	0.025	0.14,0.138	0.021	0.15,0.138	0.202
V426(β_3)		0.54,0.396	0.58,0.241*	0.008	0.58,0.241*	0.015	0.56,0.242*	0.009	0.60,0.245*	0.060
v525(β_1)	0%	-0.08,0.049	-0.07,0.030*	0.011	-0.07,0.030*	0.013	-0.07,0.030*	0.013	-0.07,0.030*	0.006
v3a00z(β_1)	36%	-0.20,0.339*	-0.40,0.217	0.122	-0.36,0.209	0.117	-0.39,0.213	0.077	-0.47,0.220*	0.134
V446 (β_1)	1%	0.001,0.040	0.004,0.023	0.008	0.01,0.023	0.009	0.01,0.024	0.009	0.004,0.023	0.017
v741(β_1)	17%	0.07,0.220	-0.03,0.138	0.022	-0.05,0.139	0.039	-0.04,0.138	0.029	-0.06,0.148	0.160
v741(β_2)		-0.46,0.433	-0.19,0.249	0.062	-0.15,0.242	0.053	-0.16,0.253	0.116	-0.14,0.256	0.136
v741(β_3)		0.13,1.018	0.42,0.518	0.157	0.40,0.526	0.138	0.38,0.515	0.137	0.35,0.511	0.045
v604(β_1)	31%	-0.10,0.307	0.24,0.212	0.267	0.28,0.199	0.174	0.29,0.198	0.137	-0.24,0.237	0.225
v604(β_2)		-0.26,0.318*	-0.04,0.204	0.112	0.01,0.203	0.110	0.02,0.208	0.108	-0.37,0.233	0.117
v604(β_3)		-1.45,1.046*	-1.41,0.729	0.010	-1.17,0.809	0.296	-1.20,0.737	0.059	-1.40,0.710	0.254
v604(β_4)		-0.09,0.750	0.12,0.324	0.477	0.23,0.390	0.577	0.25,0.274	0.296	0.04,0.450	0.139
v213(β_1)	0%	0.60,0.250*	0.37,0.177*	0.009	0.37,0.177*	0.006	0.37,0.176*	0.004	0.42,0.180*	0.010

5.9.10 Univariate Cox-PH regression analysis [Rubin rules]

The study decided to fit the univariate Cox-PH model after the multivariate Cox-PH model led to invalid conclusions on the comparison of complete cases vs. imputed data as a result of the reduction in sample size. Table 15 shows that the model estimates and the corresponding standard errors for the fully observed covariates remained unchanged after multiple imputation in all the four imputation strategies. For partially observed covariates with low missing data, the estimates, standard errors and the probability values of the covariates remained similar with the estimates from the observed values. However, for the partially observed covariates with high missing data, random forests imputation strategy produced parameter estimates that were closely related to the complete case compared to the other three imputation strategies. There were no larger differences in the estimates arising from the imputation strategies that treated clusters as class variables compared

to the MICE Flat imputation. The proportion of relative increase in variance (R) due non-response was zero for all covariates with below 1% missing data, it however increased with increase in missing data. Even though parameter estimates are closely related on most of the covariates in the four imputation strategies, there are observed variations in R based on the imputation strategy.

Table 15. Univariate Cox regression analysis of imputed data from the four imputation strategies [Rubin's analysis]

Covariate	Missing(%)	Complete Case	MICE Flat		MICE HH		MICE EA		MICE RF	
		Est,SE	Est,SE	R	Est,SE	R	Est,SE	R	Est,SE	R
v511(β_1)	5%	-0.05,0.020*	-0.05,0.019*	0.014	-0.05,0.019*	0.023	-0.05,0.019*	0.023	-0.05,0.019*	0.028
v626(β_1)	0.02%	-0.22,0.161	-0.22,0.162	0.000	-0.22,0.162	0.000	-0.22,0.162	0.000	-0.22,0.162	0.000
v626(β_2)		0.23,0.150	0.23,0.150	0.000	0.23,0.150	0.000	0.23,0.150	0.000	0.23,0.150	0.000
V426(β_1)	1%	0.04,0.269	0.03,0.269	0.005	0.03,0.269	0.020	0.02,0.269	0.020	0.04,0.269	0.022
V426(β_2)		0.22,0.136	0.23,0.134	0.011	0.21,0.135	0.018	0.21,0.136	0.034	0.22,0.135	0.017
V426(β_3)		0.69,0.240*	0.69,0.239*	0.009	0.69,0.239*	0.016	0.66,0.239*	0.004	0.70,0.242*	0.056
v525(β_1)	0%	-0.09,0.025*	-0.09,0.025*	0.000	-0.09,0.025*	0.000	-0.09,0.025*	0.000	-0.09,0.025*	0.000
v3a00z(β_1)	36%	-0.59,0.215*	-0.21,0.363	0.782	-0.36,0.282	0.563	-0.14,0.370	0.810	-0.60,0.218*	0.141
V446 (β_1)	1%	-0.01,0.022	-0.01,0.022	0.010	-0.01,0.022	0.017	-0.01,0.022	0.013	-0.01,0.022	0.008
v741(β_1)	17%	-0.20,0.140	-0.21,0.131	0.024	-0.22,0.133	0.048	-0.21,0.132	0.026	-0.21,0.142	0.170
v741(β_2)		-0.20,0.248	-0.23,0.247	0.058	-0.18,0.241	0.054	-0.20,0.252	0.113	-0.18,0.255	0.138
v741(β_3)		-0.37,0.509	0.34,0.516	0.155	0.27,0.517	0.116	-0.31,0.514	0.142	-0.31,0.511	0.049
v604(β_1)	31%	-0.10,0.222	0.16,0.205	0.247	0.20,0.196	0.177	0.23,0.195	0.107	-0.26,0.221	0.211
v604(β_2)		-0.43,0.218*	-0.11,0.196	0.093	-0.07,0.196	0.101	-0.05,0.201	0.093	-0.38,0.213	0.089
v604(β_3)		-1.88,0.731*	-1.58,0.725*	0.007	-1.35,0.110*	0.304	-1.47,0.733*	0.057	-1.52,0.700*	0.250
v604(β_4)		-0.13,0.449	0.03,0.333	0.524	0.14,0.371	0.549	0.20,0.266	0.283	-0.073,0.446	0.136
v213(β_1)	0%	0.44,0.165*	0.44,0.165*	0.000	0.44,0.165*	0.000	0.44,0.165*	0.000	0.44,0.165*	0.000

6 Discussion and conclusions

6.1 Discussion

The present study set out to solve the problem of under-five child survival missing covariate data in the Demographic and Health Survey (DHS) data sets. We used the data from the 2015-16 Tanzania DHS data to conduct risk-factor selection and perform multiple imputation. Our study findings have shown that multiple imputation can potentially solve the problem of missing data whenever studying the risk-factors of under-five child survival. Our study findings are in support of prior studies that found multiple imputation to be a plausible strategy for handling missing data in survival settings (Van Buuren et al., 1999; Eisemann et al., 2011). The study findings from the convergence plots suggest that imputations under the random forests strategy achieved better convergences compared to other imputation strategies. This finding is in contrast with a prior finding that found random forests to have achieved poor converges compared to predictive mean matching and polytomous regression in a cancer study (Eisemann et al., 2011). Bureen (2001) (Buuren and Groothuis-Oudshoorn, 2011) points out that non-convergence may indicate a problem with the variable and this should be investigated further. Our investigations suggest that two of the variables with non-convergence in the four strategies had utmost two observations with missing data. However, our results also shown that some variables converged for some imputation strategies and didn't converge for other strategies indicating that sometimes the choice of imputation strategy may affect the convergence process. Our study further indicated that that random forests imputation strategy achieved more closely identical marginal distributions for imputed vs. observed values compared to the other imputation strategies. This result suggests that random forests approach has the potential to produce imputations that are more closely related to the observed values compared to the three MICE strategies. Prior studies reviewed focused on reporting of the parameter estimates and bias arising from the imputations while ignoring the convergence or the marginal distribution of the observed vs. imputed values. Our study findings suggested that covariates that achieved both healthy convergences and completely identical marginal distributions produced good model estimates. This is an area that needs to be emphasized in reporting for future research.

The present study results also indicated that treating the two cluster variables as classes in the imputation model had no observed effects on the findings. This result suggests that there were either no big variations in the observations within the clusters or treating clusters as classes in the imputation model without adjusting the imputation method may not necessarily mean that within cluster imputation will be done. The two cluster variables

were also not ranked among the among the first 100 predictors of under-five child mortality confirming the insignificance of the clusters in the study's data set. The study findings further reveal that in the presence of low missing data, the four multiple imputation strategies produced similar results and were more closely similar to the estimates from the observed values. However, in the presence of high level of missingness, random forests shown much potential to achieve parameter estimates closely related to the estimates from the observed values. Previous studies have shown mixed results on the performance of random forests in comparison to other imputation strategies (Eisemann et al., 2011; Shah et al., 2014). Our study results also indicated that in the presence of outliers in the covariates, random forests were more likely to produce better imputations than the predictive mean matching strategy. Prior reviews have shown random forests to possess the ability to model any complexities in the data(Doove et al., 2014).

Our study findings from the random survival forests model shown that multiple imputation led to the increase in the variable importance scores and as a result all the covariates became highly predictive after multiple imputation. Even though the importance scores significantly increased, the ranking of the covariates in order of importance showed minimal discrepancies comparable to the ranking obtained using the original data relying on the RSF inbuilt missing data technique. Random survival forests imputes missing data by drawing a random value from the distribution of the observed values for the variable missing for each tree before splitting the parent node. Imputed data is however not used in the daughter nodes and ensemble cumulative hazard estimates (Ishwaran et al., 2008). Further research is needed to assess the missing data approach employed in RSF in-comparison to multiple imputation strategies. Our results also indicated that the variable importance and ranking from the four imputation strategies were closely similar, if not similar. This finding suggests that any of the imputation strategies could potentially be used to impute data as long as the objective of the study was risk factor prediction. The study findings further revealed that multiple imputation greatly affected the proportional hazards assumption in the data including the fully observed variables that were included in the imputation model. This finding suggests that much as multiple imputation doesn't alter the observed values, it may affect the original data structure. This concern has also been reported elsewhere (Pedersen et al., 2017). However, our results from the univariate cox-regression model that was fit based on the Rubin (Rubin, 1976) analysis rules suggest that multiple imputation never altered the effect of the risk-factors on the response. This finding further highlights the potential of multiple imputation in handling missing data in survival settings.
The present study shown that the under-five risk factors that were highly ranked compared well with the risk factors obtained in prior reviews. Some of the highly ranked under-five risk factors that have been found to be significant predictors of under-five mortality in earlier studies included succeeding birth (Susuman et al., 2016), total number of children ever born, number of under five children in a household, number of under three children in a household (Nasejje and Mwambi, 2017). Our study results have shown covariates related to breastfeeding were highly predictive of under-five child survival. The different country

area zones in Tanzania have also been found to be highly predictive of under-five child survival. Modeling studies may be used to assess the effect of these covariates on child mortality in Tanzania. Finally, random survival forests have shown potential to perform well in the presence of many covariates, albeit the procedure was computationally time consuming. Multiple imputation with random forests has also been found to be found to be computationally expensive in time compared to the rest of the imputation strategies.

6.2 Study strengths and limitations

Our study applied highly predictive models to identify risk factors of under-five mortality from a pool of over 400 covariates and uses multiple imputation, a flexible missing data approach that incorporates imputation uncertainty by pooling results from the multiple imputed data. Our study results may not be used to draw definitive conclusions since we don't use a simulation framework to validate them.

6.3 Future research

Our study has considered multiple imputation while treating clusters as class variables. Future research should consider multilevel imputation using clusters as levels. There are however methodological challenges with this approach especially with factor variables with more-than two levels. Future research should also consider investigating the proposed imputation strategies using the simulation frameworks and conduct sensitivity analysis as well.

Extensions of the cox proportional hazards model such as extended cox model and stratified cox model may be helpful in overcoming the proportional hazards assumption.

6.4 Recommendations

Our study suggests that most of the missing data in DHS data is actually due to skips in the questionnaire. DHS may need to start using specific codes at data collection or cleaning stages to represent skip patterns rather than treating data as missing in their online published data sets. This will also help the users of DHS data sets in knowing the true missing values.

6.5 Conclusion

Multiple imputation has shown potential to produce estimates for studying under-five child survival that are closely similar to the true estimates even in the presence of high missing data. Random forests imputation strategy shown potential to perform better than the three mice imputation strategies in achieving healthy convergences, closely similar marginal

distributions of imputed vs. observed data and model estimates. The current study results may need to be validated using a more robust simulation study and other non-response models for decisive conclusions to be made.

References

Abu, I. N., Madu, I. A., and Ajaero, C. K. (2015). The prevalence and determinants of under-five mortality in benue state, nigeria. *SAGE Open*, 5(4):2158244015611938.

Allison, P. D. (2001). *Missing data*, volume 136. Sage publications.

Allison, P. D. (2012). Handling missing data by maximum likelihood. In *SAS global forum*, volume 23. Statistical Horizons Haverford, PA, USA.

Andridge, R. R. and Little, R. J. (2010). A review of hot deck imputation for survey non-response. *International statistical review*, 78(1):40–64.

Armstrong Schellenberg, J. R., Nathan, R., Abdulla, S., Mukasa, O., Marchant, T. J., Tanner, M., and Lengeler, C. (2002). Risk factors for child mortality in rural tanzania. *Tropical Medicine & International Health*, 7(6):506–511.

Azur, M. J., Stuart, E. A., Frangakis, C., and Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research*, 20(1):40–49.

Beaujean, A. (2012). Bayloredpsych: R package for baylor university educational psychology quantitative courses. *R package version 0.5, URL http://CRAN. R-project. org/package= BaylorEdPsych*.

Bennett, D. A. (2001). How can i deal with missing data in my study? *Australian and New Zealand journal of public health*, 25(5):464–469.

Borman, S. (2004). The expectation maximization algorithm-a short tutorial. *Submitted for publication*, pages 1–9.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45(3).

Chen, X. and Ishwaran, H. (2012). Random forests for genomic data analysis. *Genomics*, 99(6):323–329.

Cox, D. R. (1972). Regression models and life-tables. In *Breakthroughs in statistics*, pages 527–541. Springer.

Dong, Y. and Peng, C.-Y. J. (2013). Principled missing data methods for researchers. *SpringerPlus*, 2(1):222.

Doove, L. L., Van Buuren, S., and Dusseldorp, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics & Data Analysis*, 72:92–104.

Ehrlinger, J. (2016). ggrandomforests: Exploring random forest survival. *arXiv preprint arXiv:1612.08974*.

Eisemann, N., Waldmann, A., and Katalinic, A. (2011). Imputation of missing values of tumour stage in population-based cancer registration. *BMC medical research methodology*, 11(1):129.

Ezeh, O. K., Agho, K. E., Dibley, M. J., Hall, J. J., and Page, A. N. (2015). Risk factors for postneonatal, infant, child and under-5 mortality in nigeria: a pooled cross-sectional analysis. *BMJ open*, 5(3):e006779.

Geman, S. and Geman, D. (1987). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. In *Readings in Computer Vision*, pages 564–584. Elsevier.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual review of psychology*, 60:549–576.

Grund, S., Lüdtke, O., and Robitzsch, A. (2017). Multiple imputation of missing data for multilevel models: Simulations and recommendations. *Organizational Research Methods*, page 1094428117703686.

Hamidi, O., Poorolajal, J., Farhadian, M., and Tapak, L. (2016). Identifying important risk factors for survival in kidney graft failure patients using random survival forests. *Iranian journal of public health*, 45(1):27.

Hsich, E., Gorodeski, E. Z., Blackstone, E. H., Ishwaran, H., and Lauer, M. S. (2011). Identifying important risk factors for survival in patient with systolic heart failure using random survival forests. *Circulation: Cardiovascular Quality and Outcomes*, 4(1):39–45.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008). Random survival forests. *The annals of applied statistics*, pages 841–860.

Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., and Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial intelligence in medicine*, 50(2):105–115.

Kleinbaum, D. G. and Klein, M. (2010). *Survival analysis*, volume 3. Springer.

Kozuki, N. and Walker, N. (2013). Exploring the association between short/long preceding birth intervals and child mortality: using reference birth interval children of the same mother as comparison. *BMC public health*, 13(3):S6.

Larsen, R. (2011). Missing data imputation versus full information maximum likelihood with second-level dependencies. *Structural Equation Modeling: A Multidisciplinary Journal*, 18(4):649–662.

Little, R. J. and Rubin, D. B. (1989). The analysis of social science data with missing values. *Sociological Methods & Research*, 18(2-3):292–326.

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

Ma, J., Thabane, L., Dolovich, L., and Akhtar-Danesh, N. (2011). Imputation strategies for missing binary outcomes in cluster randomized trials. *BMC medical research methodology*, 11(1):18.

Marshall, A., Altman, D. G., Royston, P., and Holder, R. L. (2009). Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. *BMC medical research methodology*, 9(1):57.

Masanja, H., de Savigny, D., Smithson, P., Schellenberg, J., John, T., Mbuya, C., Upunda, G., Boerma, T., Victora, C., Smith, T., et al. (2008). Child survival gains in tanzania: analysis of data from demographic and health surveys. *The Lancet*, 371(9620):1276–1283.

Mosley, W. H. and Chen, L. C. (1984). An analytical framework for the study of child survival in developing countries. *Population and development review*, 10(0):25–45.

Nasejje, J. B. and Mwambi, H. (2017). Application of random survival forests in understanding the determinants of under-five child mortality in uganda in the presence of covariates that satisfy the proportional and non-proportional hazards assumption. *BMC Research Notes*, 10(1):459.

Nasejje, J. B., Mwambi, H. G., and Achia, T. N. (2015). Understanding the determinants of under-five child mortality in uganda including the estimation of unobserved household and community effects using both frequentist and bayesian survival analysis approaches. *BMC public health*, 15(1):1003.

Pedersen, A. B., Mikkelsen, E. M., Cronin-Fenton, D., Kristensen, N. R., Pham, T. M., Pedersen, L., and Petersen, I. (2017). Missing data and multiple imputation in clinical epidemiological research. *Clinical Epidemiology*, 9:157.

Peugh, J. L. and Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of educational research*, 74(4):525–556.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Rubin, D. B. (1987). The calculation of posterior distributions by data augmentation: Comment: A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The sir algorithm. *Journal of the American Statistical Association*, 82(398):543–546.

Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical methods in medical research*, 8(1):3–15.

Schafer, J. L. and Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological methods*, 7(2):147.

Schafer, J. L. and Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate behavioral research*, 33(4):545–571.

Schenker, N. and Taylor, J. M. (1996). Partially parametric techniques for multiple imputation. *Computational statistics & data analysis*, 22(4):425–446.

Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., and Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using mice: a caliber study. *American journal of epidemiology*, 179(6):764–774.

Soullier, N., de La Rochebrochard, E., and Bouyer, J. (2010). Multiple imputation for estimation of an occurrence rate in cohorts with attrition and discrete follow-up time points: a simulation study. *BMC medical research methodology*, 10(1):79.

Sun, J. (2007). *The statistical analysis of interval-censored failure time data*. Springer Science & Business Media.

Susuman, A. S. and Hamisi, H. F. (2012). Under-5 mortality in tanzania: A demographic scenario. *Iranian journal of public health*, 41(12):8.

Susuman, A. S., Hamisi, H. F., and Nagarajan, R. (2016). Bio-demographic factors affecting child loss in tanzania. *Genus*, 72(1):10.

Templ, M., Alfons, A., Kowarik, A., and Prantner, B. (2011). Vim: visualization and imputation of missing values. *R package version*, 2(3).

UNDESA (2015). World population prospects: The 2015 revision, key findings and advance tables. Technical report, United Nations Department of Economic and Social Affairs and Population Division.

Unicef (2015). *Levels & Trends in Child Mortality: Report 2015: Estimates Developed by the UN Inter-Agency Group for Child Mortality Estimation*. United Nations Children's Fund.

Van Buuren, S., Boshuizen, H. C., Knook, D. L., et al. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in medicine*, 18(6):681–694.

Vink, G., Frank, L. E., Pannekoek, J., and Buuren, S. (2014). Predictive mean matching imputation of semicontinuous variables. *Statistica Neerlandica*, 68(1):61–90.

Yaya, S., Ekholuenetale, M., Tudeme, G., Vaibhav, S., Bishwajit, G., and Kadio, B. (2017). Prevalence and determinants of childhood mortality in nigeria. *BMC public health*, 17(1):485.