BMJ Open

Use of a deep learning and random forest approach to track changes in the predictive nature of socioeconomic drivers of under-5 mortality rates in sub-Saharan Africa

Justine B Nasejje ,[1] Rendani Mbuvha,[1] Henry Mwambi[2]

To cite: Nasejje JB, Mbuvha R, Mwambi H. Use of a deep learning and random forest approach to track changes in the predictive nature of socioeconomic drivers of under-5 mortality rates in sub-Saharan Africa. *BMJ Open* 2022;**12**:e049786. doi:10.1136/bmjopen-2021-049786

► Prepublication history for this paper is available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2021-049786).

Received 02 February 2021
Accepted 13 January 2022

[1]Statistics and Actuarial Science, University of the Witwatersrand, Johannesburg-Braamfontein, South Africa
[2]School of Mathematics, Statistics and Computer Science, University of Kwazulu-Natal, Pietermaritzburg, South Africa

Correspondence to
Dr Justine B Nasejje;
justine.nasejje@wits.ac.za

ABSTRACT

Objectives We used machine learning algorithms to track how the ranks of importance and the survival outcome of four socioeconomic determinants (place of residence, mother's level of education, wealth index and sex of the child) of under-5 mortality rate (U5MR) in sub-Saharan Africa have evolved.

Settings This work consists of multiple cross-sectional studies. We analysed data from the Demographic Health Surveys (DHS) collected from four countries; Uganda, Zimbabwe, Chad and Ghana, each randomly selected from the four subregions of sub-Saharan Africa.

Participants Each country has multiple DHS datasets and a total of 11 datasets were selected for analysis. A total of n=85 688 children were drawn from the eleven datasets.

Primary and secondary outcomes The primary outcome variable is U5MR; the secondary outcomes were to obtain the ranks of importance of the four socioeconomic factors over time and to compare the two machine learning models, the random survival forest (RSF) and the deep survival neural network (DeepSurv) in predicting U5MR.

Results Mother's education level ranked first in five datasets. Wealth index ranked first in three, place of residence ranked first in two and sex of the child ranked last in most of the datasets. The four factors showed a favourable survival outcome over time, confirming that past interventions targeting these factors are yielding positive results. The DeepSurv model has a higher predictive performance with mean concordance indexes (between 67% and 80%), above 50% compared with the RSF model.

Conclusions The study reveals that children under the age of 5 in sub-Saharan Africa have favourable survival outcomes associated with the four socioeconomic factors over time. It also shows that deep survival neural network models are efficient in predicting U5MR and should, therefore, be used in the big data era to draft evidence-based policies to achieve the third sustainable development goal.

Strengths and limitations of this study

- The study used machine learning methods which when compared with classical statistical models are very flexible.
- Machine learning methods have fewer assumptions and are adapted to fit very large datasets with complex relations between predictors and a given outcome.
- Machine learning models may not give an effect size of the factors.
- With these methods, it is very difficult to tell by how much the factor affects the outcome.
- Causes of death of the children were unknown at the time of the survey.

INTRODUCTION

Reducing under-5 mortality rate (U5MR) was the fourth of the Millennium Development Goals (MDGs) drafted in the year 2000, and the world sprang into action to achieve it, and it now appears within the third Sustainable Development Goal (SDG3).

The probability of a child dying before the age of 5 is a global indicator of societal and national development; it serves as a key marker of health equity and access.[1] The fourth MDG (MDG4), which centred at reducing under-5 mortality by two-thirds in the period between 1990 and 2015, now appears in the third SDG (SDG3). It is to 'Ensure healthy lives and promote well-being for all at all ages'. Although U5MR has declined in most sub-Saharan countries, there are substantial inequalities that still exist between subgroups of the population within countries.[2 3] These subgroups are based on factors such as: wealth index, maternal factors such as education level, place of residence and the sex of the child, among others. The Mosley and Chen framework categorises these socioeconomic factors as the distal determinants of child mortality.[4]

Classical statistical parametric regression models such as the logistic regression model,

semiparametric models like the Cox proportional hazard model (CPH), and generalised additive models, have been widely used to study determinants of U5MR.[1 5–11] Sahu *et al*,[7] study on levels, trends and predictors of infant and child mortality among tribes in rural India, used the CPH model to understand the socioeconomic and demographic factors associated with mortality from 1992 to 2006 in India. The study concluded that household wealth is significantly associated with infant and child mortality. They also concluded that mortality differentials by sociodemographic and economic factors were observed over the period. Mother's education level and sex of the child were among the factors responsible for the trends and differentials of U5MR in rural India. Similar studies in Nigeria concluded that place of residence (rural or urban) was an important risk factor in determining U5MR along with mother's education and sex of the child.[12 13] Although the CPH and the logistic regression models are very robust, they are often criticised for their restrictive assumptions and potentially lead to bias if one does not take care when preparing data for analysis.[14] Classical machine learning approaches which include nearest neighbours, neural networks, kernel methods, penalised least squares and data partitioning methods, such as decision trees (CART) and random forests, are among the alternative approaches to parametric and semiparametric classical models.[15–17] Recently, deep learning methods, which are advances in neural networks, have been recommended for analysing survival data.[18–24] These machine learning models are known to be very flexible compared with the statistical models like the CPH model.[21–25] A recent study by Adegbosin *et al*,[25] recommended using deep learning models to understand the determinants of U5MR in low-income and middle-income countries.

Previous studies have shown that the four socioeconomic factors; place of residence, mother's education, household wealth index and sex of the child, are often stated among the top predictors of under-5 mortality in the sub-Saharan region.[12–25] With the launch of the MDG in the year 2000, we saw the convergence of the development agendas of United Nations Development Programme; United Nations Environment Programme; WHO; UNICEF; UNESCO and other development agencies, to raise funding and create programmes to combat existing inequalities to achieve these goals.[26] Despite the substantial improvement made with the MDG4, inequalities persist today, and progress has been uneven. Now that the MDG4 appears as a facet of the SDG3 with an even wider age range, we need an evidence-based approach to achieve it by using existing datasets to inform policy.

Studying how the rank in importance of these factors to determine U5MR has evolved over time can help redirect resources to the right sectors, and hence be on-course to achieve SDG3. In this study, therefore, we train a random survival forest (RSF) and deep survival neural network model to understand how the rank of importance, the survival outcome and predictive nature of these socioeconomic factors in determining U5MR in sub-Saharan Africa have evolved over time. The RSF model is used to rank importance of these factors. The deep survival neural network model is used to determine whether these factors are still predictive, and to extract survival curves to assess whether there is a favourable survival outcome for children under the age of 5 associated with these factors in this region over time.

The contributions of this work are as follows: (1) to identify the rankings of the four socioeconomic factors in U5MR prediction in sub-Saharan Africa; (2) to present how the ranking of these factors has changed over time and (3) to present an application of deep survival models in modelling U5MR in the sub-Saharan Africa region to identify changes in the survival outcome associated with the four economic factors. These contributions are aimed at assisting policymakers in designing new interventions and providing evidence of how past interventions have worked through presenting changes in predictive importance rankings of the four socioeconomic factors over time.

METHODS

This study uses two machine learning models; the RSF model, and the deep survival neural network to answer the following questions: What are the ranks of importance of the four social socioeconomic factors over time for countries in the sub-Saharan region? Are the four socioeconomic factors linked to a favourable survival outcome in the region over time, especially after the expiry of the MDGs? Which of the two machine learning methods, the RSF and the DeepSurv model, is effective in predicting U5MR?

Data

Eleven datasets of completed Standard Demographic and Health Surveys (DHS) from four countries in sub-Saharan Africa were used for this study. The four countries were randomly selected from the four subregions (Southern, Central, Eastern and Western Africa) of sub-Saharan Africa. DHS is funded by USAID, UNFPA, UNICEF, Irish Aid and the government of the UK and since 1988 has provided datasets rich in information on fertility, family planning, maternal and child health, gender, HIV/AIDS, malaria and nutrition in sub-Saharan Africa. The survey uses a two-stage cluster sampling.[25] More information about the sampling design, data collection and processing details are described on the DHS programme website. The datasets are available on request from the DHS programme. The outcome variable is under-5 survival time, and this information was obtained from the birth history of interviewed women aged from 15 to 49 years. All datasets used in this analysis are comprised of both living and deceased children, born in the period of 5 years preceding the date of the survey. This is to limit the gap between the event and collection of socioeconomic information. The socioeconomic factors in this study were restricted to place of residence, mother's level of

Table 1 The standard DHS datasets used for this study, by subregions of sub-Saharan Africa identified by the year the survey was conducted

Southern region	Eastern region
Zimbabwe	Uganda
2006	2006
2011	2011
2015	2016
Western region	**Central region**
Ghana	Chad
2003	
2008	2004
2014	2014

DHS, Demographic Health Surveys.

education, wealth index of the household and sex of the child. The four countries and the DHS datasets selected from each subregion are shown in table 1.

Data preprocessing

DHS datasets contain many features or variables. In this study only four features were considered for analysis: place of residence, mothers' level of education, wealth index and sex of the child. Other features were excluded. The outcome variable, survival time, was calculated differently, depending on the survival status of the child. Children under the age of 5 that were living at the time of the survey had their survival time calculated as the difference between the year of the interview and year of birth. For children who were deceased at the time of the survey, survival time was calculated as the difference between the year of the interview and the year of death. Survival time was measured in months for this analysis. For each dataset, a data frame containing the four features, survival time and the status indicator (living or deceased), was created. While information was complete across all datasets for the features considered in this analysis, some of the datasets that were collected in the 1990s and the early 2000s, wealth index was not a recorded feature. These datasets were excluded in our final analysis to allow meaningful comparisons. Tables 2 and 3 give the counts of the number of children under the age of 5 for each of the feature category in all the datasets considered for analysis.

Patient and public involvement

There were no patients involved in this study.

Models

The CPH model is the most prominent model for analysing survival data.[1 5] However, its assumption that the outcome (log hazard) is a linear combination of the covariates, is too restrictive to predict survival outcomes which are complex and involve higher interactions between predictive variables. This creates the need to use models that are more flexible in predicting survival outcomes. Classical machine learning techniques, such as survival trees and RSF, enable the detection of complex relationships in survival datasets, and they have been employed in recent years.[15] These methods have achieved high accuracy in predicting the survival outcomes when applied to survival datasets to identify factors affecting U5MR.[27]

Table 2 Number of children under by sex of the child, place of residence and mother's education level

	Sex of The child		Place of Residence		Mother's education Level					
	Male	Female	Urban	Rural	None	Incomplete Primary	Complete Secondary	Incomplete Secondary	Complete Secondary	Higher
Zimbabwe										
2006	2636	2610	1340	3906	206	1696	330	2870	22	122
2011	2812	2751	1611	3952	100	710	1131	3417	54	151
2015	3024	3108	2316	3816	63	736	1070	3823	78	362
Uganda										
2006	4145	4224	917	7452	2034	4346	835	932	27	195
2011	3944	3934	1682	6196	1427	3789	898	1361	84	319
2016	7844	7678	2811	12711	2080	7568	2137	2767	162	808
Chad										
2004	2839	2796	2504	3131	4174	943	119	341	29	29
2014	9472	9151	3973	14650	13424	2898	730	1329	165	77
Ghana										
2003	1950	1894	1043	2801	1824	595	228	1069	88	40
2008	1526	1466	1000	1992	1132	561	161	924	149	65
2014	3066	2818	2344	3540	2042	884	325	2055	354	224

Table 3 Number of children under 5 by wealth index

	Wealth index					
	Poorest	Poorer	Middle	Richer	Richest	Total
Zimbabwe						
2006	1351	1166	958	1019	752	**5246**
2011	1366	1145	1001	1178	873	**5563**
2015	1244	1075	958	1603	1252	**6132**
Uganda						
2006	2139	1820	1555	1491	1364	**8369**
2011	2030	1550	1405	1230	1663	**7878**
2016	4152	3382	2971	2607	2410	15522
Chad						
2004	916	867	762	1011	2079	**5635**
2014	3559	3786	3902	4097	3279	18623
Ghana						
2003	1285	859	682	539	479	**3844**
2008	973	656	504	502	357	**2992**
2014	1886	1304	1083	883	728	**5884**

The total number of children from all the DHS datasets used in this study is 85688.
DHS, Demographic Health Surveys.

Even though they have exhibited a good performance in predicting survival outcomes, there are few studies aimed at understanding factors associated with U5MR that have embraced these methods.[15 27] Recently, with the advancement of machine learning methods, deep learning methods have also been added to the toolbox of methods to analyse survival data.[21] Because most datasets collected have complex structures, using models that have very strict assumptions, may lead to bias, thus misleading policy implementations. In this study, we applied two machine learning models on datasets from sub-Saharan Africa. They are the RSF, and the deep survival neural network model (DeepSurv).[17 21]

Random survival forests

RSFs are an extension of regression trees formally presented by Breiman *et al*,[28] to survival data. These methods have been found to be the most desirable in addressing the challenges of the CPH model. First, we describe the survival tree, an important building block of the forest. This is followed by the algorithm of the RSF model by Breiman *et al*.[28]

Survival trees

The regression tree algorithm for right censored data, is an extension of the Classification and Regression Trees (CART) algorithm by Breiman *et al*.[28] Box 1 is the general algorithm for survival trees.[29–31]

An RSF model is a collection of survival trees because a single tree is not always a good probability estimator due to its shortcomings of giving unstable estimators.[32 33] Researchers have, over the years, recommended the growing of an entire forest as the solution to the shortcomings of a single tree. Box 2 for building an RSF model as presented by Ishwaran *et al*,[17] is given below as follows:

Note that the node size is restricted such that the number of unique events at a node does not drop below the minimum number.

This study used a special type of survival forest model known as the conditional inference survival forest model (CIF).[34 35] The CIF has the advantage, over the original RSF algorithm, of correcting the bias that results from favouring covariates that have many split points, rather than choosing covariates that are highly associated with the outcome.[15 17 35 36]

The random survival model was trained in the R-software with each forest consisting of 200 trees (Code).[37 38]

Neural network survival models

Non-linear models, like artificial neural networks, are becoming increasingly popular as additional models in the

Box 1 Algorithm 1: survival tree algorithm

1. At each node, each covariate and all its allowable split points are candidates for splitting the node into two daughter nodes.
2. Compute the impurity measure based on a predetermined split-rule at the node on a pool of all allowable split points.
3. Split the node into two daughter nodes (α and β) using the value of an impurity measure. The best split maximises the difference between the two daughter nodes.
4. Recursively repeat steps 2 and 3 by treating each daughter node as a root node.
5. Stop if a node is terminal, that is, has no less than d_0 >0 unique observed events.

Box 2 Algorithm 2: survival forest algorithm

1. Draw B, bootstrap samples from the original data set. Each bootstrap sample, b=1, 2…, B excludes about 30% of the data and this is called out-of-bag (OOB).
2. Grow a survival tree for each bootstrap sample, at each node randomly select a subset of covariates. Split the node by selecting the covariate that maximises the difference between daughter nodes using a predetermined split rule.
3. Grow the tree to full size under the constraint that a terminal node should have no less than d_0 >0 unique death.
4. Calculate the cumulative hazard ($\hat{\Lambda}(t)$) or survival curve ($\hat{S}(t)$) for each tree. Average to obtain the ensemble estimate.
5. Using OOB data, calculate prediction error for the ensemble cumulative hazard function (CHF) or survival probability.

toolbox of models aimed at predicting survival outcomes. They look very promising, especially when applied to large datasets that could have many covariates with non-linear effects on the survival outcome. It is important to note that neural networks are only prominent for predicting outcomes, but they cannot give explanations or quantify covariate effects on the outcomes. Initially, a single hidden layer feed-forward neural network was trained to survival data and its performance in predicting survival outcomes provided mixed results.[21–24] Recently, with the introduction of deep learning methods, which are advances in neural networks, deep survival neural networks have been found to gain superiority over existing methods in predicting survival outcomes.[18–20] Instead of only one hidden layer in the neural network, more than one hidden layer is used. The Neural net considered in this study is based on the likelihood function of the CPH model.[39] Therefore, before describing the neural network, we give a brief introduction to the CPH model.

CPH model

The hazard function depends on time t and a vector of covariates X through:

$$\underline{\lambda}(t, X) = \lambda_0(t)\exp\left(h(X)\right) \tag{1}$$

Where $\lambda_0(t)$ is the baseline hazard function and $\exp(h(X))$ the risk score. The CPH model estimates $h(X)$, by a linear function $h_\beta(X) = \beta' \cdot X$. The estimates $\left(\beta\right.$ of the parameters (β) are obtained by maximising the partial likelihood. Suppose that there are k distinct event times, and $t_1 < t_2 < \ldots < t_k$ represent the ordered distinct event times, the partial likelihood is given as:

$$\mathrm{L}(\beta) = \prod_{\mathrm{i}=1}^{\mathrm{k}} \mathrm{i}\frac{\exp\left(\hat{\mathrm{h}}_\beta(\mathrm{X_i})\right)}{\sum_{\mathrm{j}\in \mathrm{R}(\mathrm{t_i})}\exp\left(\hat{\mathrm{h}}_\beta(\mathrm{X_j})\right)} \tag{2}$$

This estimation of $h(X)$ by $\hat{h}_\beta(X)$ is very restrictive and can lead to biased results for studies where it is violated. This criticism has led to the need to use more flexible models to analyse survival datasets. Neural networks are among these new methods for survival analysis. A neural network consists of an input layer, hidden layers and an output layer. Each input is connected directly to all but one node in the hidden layer. A non-linear transformation is performed on a weighted sum of the inputs. The rectified linear activation function (ReLU) is recommended in modern neural networks as the transformation or activation function to compute hidden layer values. This is defined as:

$$g(z) = max\{0, z\} \tag{3}$$

In this study, however, the Scaled Exponential Linear Unit (SELU) is used as an activation function because of its advantages over the ReLU as it can get trapped in a dead state. That is, the weights' change is so high, and the resulting z in the next iteration so small such that the activation function is stuck at the left side of zero. The affected cell cannot contribute to the learning of the network anymore, and its gradient stays at zero. If this happens to numerous cells in your network, the power of the trained network stays below its theoretical capabilities. It is given as:

$$\mathrm{g}(z) = \lambda\begin{cases}\gamma\left(exp(z) - 1\right), & z < 0,\\ z, & z \geq 0.\end{cases}$$

Where γ>0 and λ>0 are to be specified and chosen such that the mean and variance of the inputs are preserved between two consecutive layers. It looks like a ReLU for values larger than zero, there is an extra parameter involved, λ. This parameter is the reason for the S(caled) in SELU. Consider replacing the linear function $h_\beta(X) = \beta' \cdot X$ in equation 2 by the output of $h_\theta(X) = \exp\left(g(X, \theta)\right)$ of the neural network. The proportional hazards model becomes

$$h_\theta(X_i) = \exp\left(g(X_i, \theta)\right) \tag{4}$$

This implies that the covariates of the uppermost hidden layer of the deep network are used as the input to the CPH model. The output of the deep neural network is a single node that contains estimates of the risk function in equation 4 $\left(\hat{h}_\theta(t, X_i)\right)$ and the function to be maximised is:

$$\mathrm{L}(\theta) = \prod_{\mathrm{i}:\delta_\mathrm{i}=1}\frac{\exp\left(\hat{\mathrm{h}}_\theta(\mathrm{X_i})\right)}{\sum_{\mathrm{j}\in \mathrm{R_i}(\mathrm{t_i})}\exp\left(\hat{\mathrm{h}}_\theta(\mathrm{X_j})\right)} \tag{5}$$

The average negative log partial likelihood of equation 5 is given as:

$$l(\theta) = -\frac{1}{n_{\delta_1}}\sum_{i:\,\delta_i=1}\left(\hat{h}_\theta(X_i) - log\sum_{j\in R(t_i)} exp\left(\hat{h}_\theta(X_j)\right)\right) \tag{6}$$

where n_{δ_1} is the number of events in the dataset. To penalise for model complexity, a term is added to the loss function to put weight on a few of the covariates. Penalty of ridge regression or L_2-norm is used in this study. The loss function to be minimised is therefore given as:

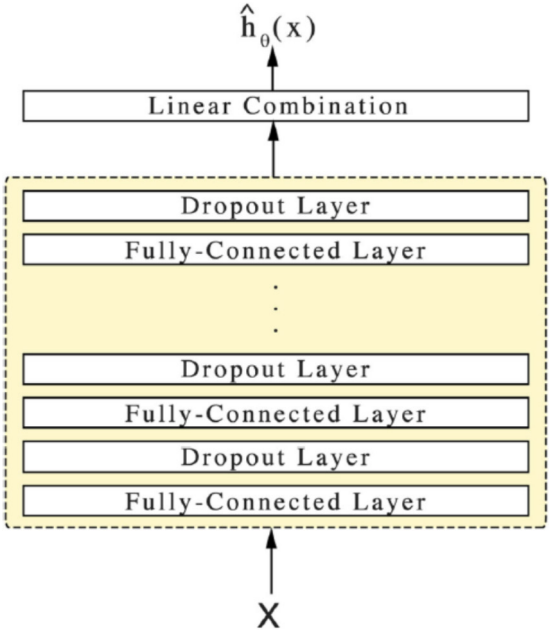

Figure 1 DeepSurv architecture Katzman *et al*.[21]

$$l(\theta) = -\frac{1}{n_{\delta_1}} \sum_{i:\,\delta_i=1} \left(\hat{h}_\theta(X_i) - log \sum_{j \in R(t_i)} exp\left(\hat{h}_\theta(X_j)\right) \right) + \alpha\theta_2^2 \quad (7)$$

Therefore, the network is trained by setting the objective function to be the average negative log partial likelihood of the CPH model with regularisation where α is the regularisation parameter for the L_2 norm. Gradient descent optimisation is used to find the weights of the network which minimise the loss function. The DeepSurv neural network architecture is described in detail by Katzman *et al*.[21] Figure 1 shows its architecture. It is a deep feed-forward neural network implemented as:

DeepSurv was popularised by Katzman *et al*,[21] who implemented it in *Theano* Python library with the Python package *Lasagne*. In this study, however, we used the PySurvival python package implementation of the same model by Fotso.[40] For our study, observed socioeconomic factors are given as inputs to the network. The hidden layers of the network consist of a fully connected layer of nodes, followed by a dropout layer. The output layer has one node with a linear activation which estimates the log-risk function in the CPH model. The loss function for the network is shown in equation 7. A drop-out probability is introduced such that at each training stage, individual nodes are either dropped out of the network with probability 1 − p or kept with probability p, so that a reduced network is left to prevent overfitting. In this study, p=0.2 and a learning rate of 1e-8 are used (Code).

Model evaluation

The Concordance index (C-index) is a common metric used to evaluate the performance of survival models. It is defined as the probability of agreement for any two randomly chosen observations, where agreement means that the observation with the shorter survival time should have the larger risk score, and the opposite is true.[41 42] Note that censored observation cannot be compared with any observed event time because its exact event time is unknown; however, any other pair of observations are called comparable.[43] If predicted survival outcomes are denoted by Y, the C-index is given by:

$$C = \frac{\sum_{i:\delta_1=1} \sum_{yi<yj} I(\hat{Y}i<\hat{Y}_j)}{\text{Number of comparable pairs}} \quad (8)$$

In survival analysis, shorter survival time means smaller predicted outcomes. C-index value of above 0.5 means better agreement among comparable pairs.[41–43] Overfitting is one of the criticisms of machine learning techniques. This arises from using the training error to evaluate the model performance. In this study, we used a cross-validated C-index to evaluate the performance of the deep learning model.

Cross-validation

Splitting the data into a test and train set is one of the most used methods to evaluate the predictive performance of machine learning models. The test error is known to be more informative than the train error, because of the assumption that the test dataset is independent from the train dataset. However, the test error can vary from one test sample to another and, since the test data is a subset of the train set, this independence is not guaranteed. This makes this method unreliable. Hence K-fold cross-validation is recommended. K-fold cross-validation divides the data into K folds and ensures that each fold is used as a testing set at some point.[44] In this study, we used a 10-fold cross-validation. The dataset is divided into 10 folds or sections. The first fold is set aside to use as a test set and the rest of the folds combine to serve as the training set. In the second iteration, the second fold is used as the testing set while the rest serve as the training set. This process is repeated until each of the ten folds have been used as the testing set.

Measures of covariate importance

To understand which factors are important in influencing predictions, the RSF model has a measure which estimates the importance of each covariate. It is generally referred to as the variable importance measure.[45–48] Variables are selected because of their importance in predicting the survival outcome. The basic measure of variable importance is to count the number of times the predictor is selected by each tree in the whole forest.[49] Different measures of variable importance exist in literature and have been implemented in the random forest algorithms.[28 32 49 50] In this study, permutation importance was selected as our measure of covariate importance.

Permutation importance

Permutation importance is based on the idea of identifying whether the covariate in question has a positive effect on the predictive performance of the random forest model. As an illustration, first consider a tree grown and its prediction accuracy (e), calculated by using

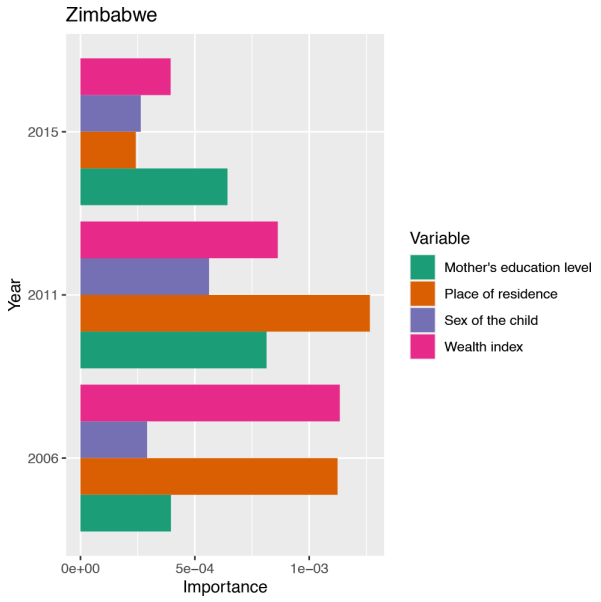

Figure 2 Ranks of importance for the four socioeconomic factors in predicting U5MR in Zimbabwe over a period of 9 years. U5MR, under-5 mortality rate.

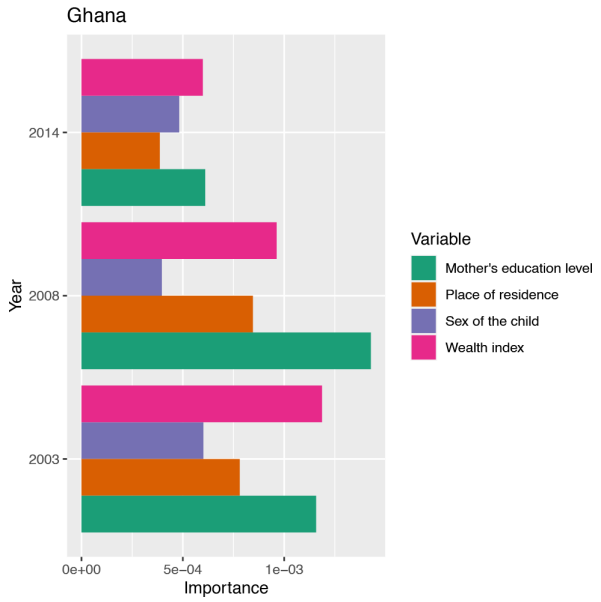

Figure 3 Ranks of importance for the four social economic factors in predicting U5MR in Ghana over a 10-year period. U5MR, under-5 mortality rate

the out-of-bag (OOB) observations. Second, randomly permute the values of the factor of interest, (X_i) for all individuals. Note that permutation breaks the original relationship of the covariate with the survival outcome. Obtain a new value for prediction accuracy, (e_i) using OOB observations. Compare e_i, with e of the original classification for covariate, X_i . Calculate, argmax {0; e_i − e }. The difference between the accuracy before and after permutation provides the importance of the covariate X_i from a single tree. Permutation variable importance of a covariate for the entire forest is calculated by averaging over all the tree importance values. This is repeated for all covariates of interest.[32 50 51]

RESULTS

In this study, we applied the random forest algorithm described in the methods section on the selected datasets, and we extracted the most important variables in predicting child survival. We used a special type of the RSF model known as the CIF model. This was done to avoid the bias that results from favouring covariates that have many split points, rather than choosing covariates that are highly associated to the outcome. The ranks of importance of the four features obtained by applying the CIF to the datasets are shown in figures 2–5. The ranks of feature importance presented here are for datasets from each country that was selected from each subregion.

In figure 2, the two most important predictors of U5MR in Zimbabwe in 2006 are wealth index and place of residence, respectively. In 2011, place of residence and wealth index are ranked as the most predictive factors of U5MR. Lastly, in 2015, mother's education and place of residence are the top ranked predictors.

In figure 3, mother's education is ranked first for the years 2008 and 2014, and wealth index second in both datasets.

In figure 4, wealth index and mother's education are ranked first and second in 2006. Wealth index and mother's education are ranked first and second in 2011. Lastly in 2016, mother's education is ranked first, and wealth index is ranked second in predicting U5MR in Uganda. Figure 5 shows that place of residence and wealth index are ranked the top two most important predictor variables in predicting U5MR in Chad.

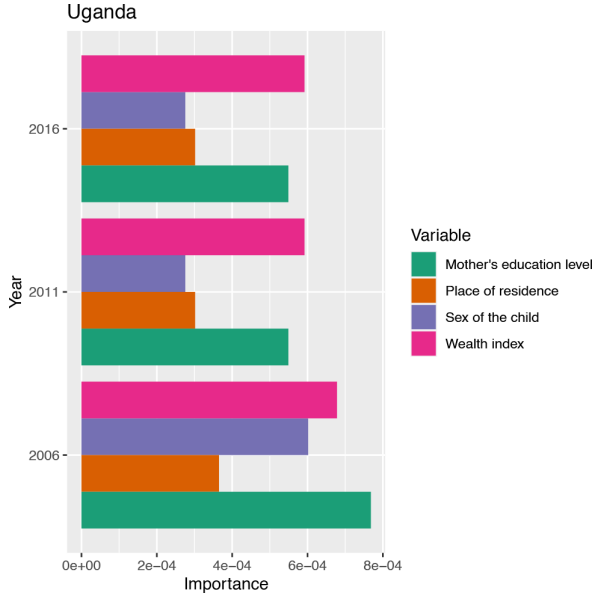

Figure 4 Ranks of importance for the four social economic factors in predicting U5MR in Uganda over a period of 10 years. U5MR, under-5 mortality rate.

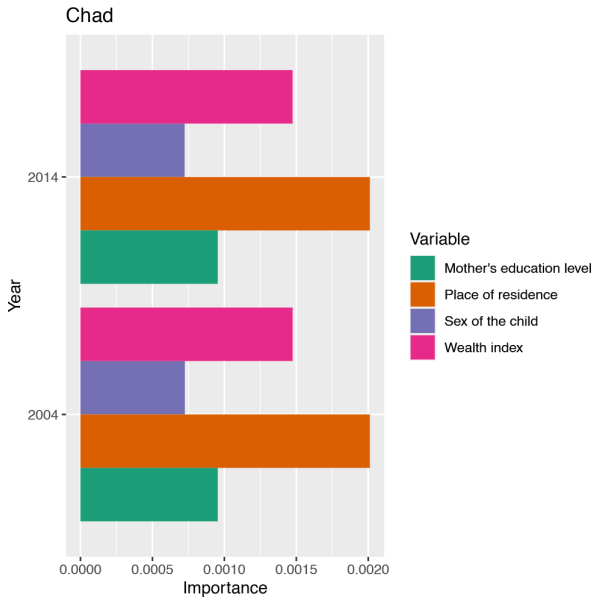

Figure 5 Ranks of importance for the four social economic factors in predicting U5MR in Chad over the period of 10 years. U5MR, under-5 mortality rate

Figures 2–5 show that mother's education is ranked first in 5 out of the 11 datasets, and wealth index ranked first in three out of the eleven datasets, but second in 8 out of the 11 datasets. This shows that these two factors are dominant in predicting U5MR in the region over time. Place of residence has also been ranked first in 2 out of the 11 datasets, and second in 1 of the 11 datasets, placing it among the top three predictors of under-5 survival in the countries considered in this study.

It is evident from these rankings that mother's education and wealth index were among the most dominant factors. The sex of the child is not anywhere near the top two ranks of importance in all the datasets considered for analysis. In fact, it was ranked last in six out of the eleven datasets.

These results agree with a study by Rutstein *et al*,[52] which studied the changes in socioeconomic inequalities in low-income and middle-income countries in the 2000s.

The study also applied the DeepSurv model to the selected datasets and extracted survival curves from the model output to establish whether the survival outcome associated with the four socioeconomic factors has become favourable over time.

Figure 6 shows survival curves of the survival outcome (under-5 survival time), associated with the four socioeconomic factors extracted from the deep learning survival model, for the test datasets obtained from the eleven datasets of the four countries from the four subregions

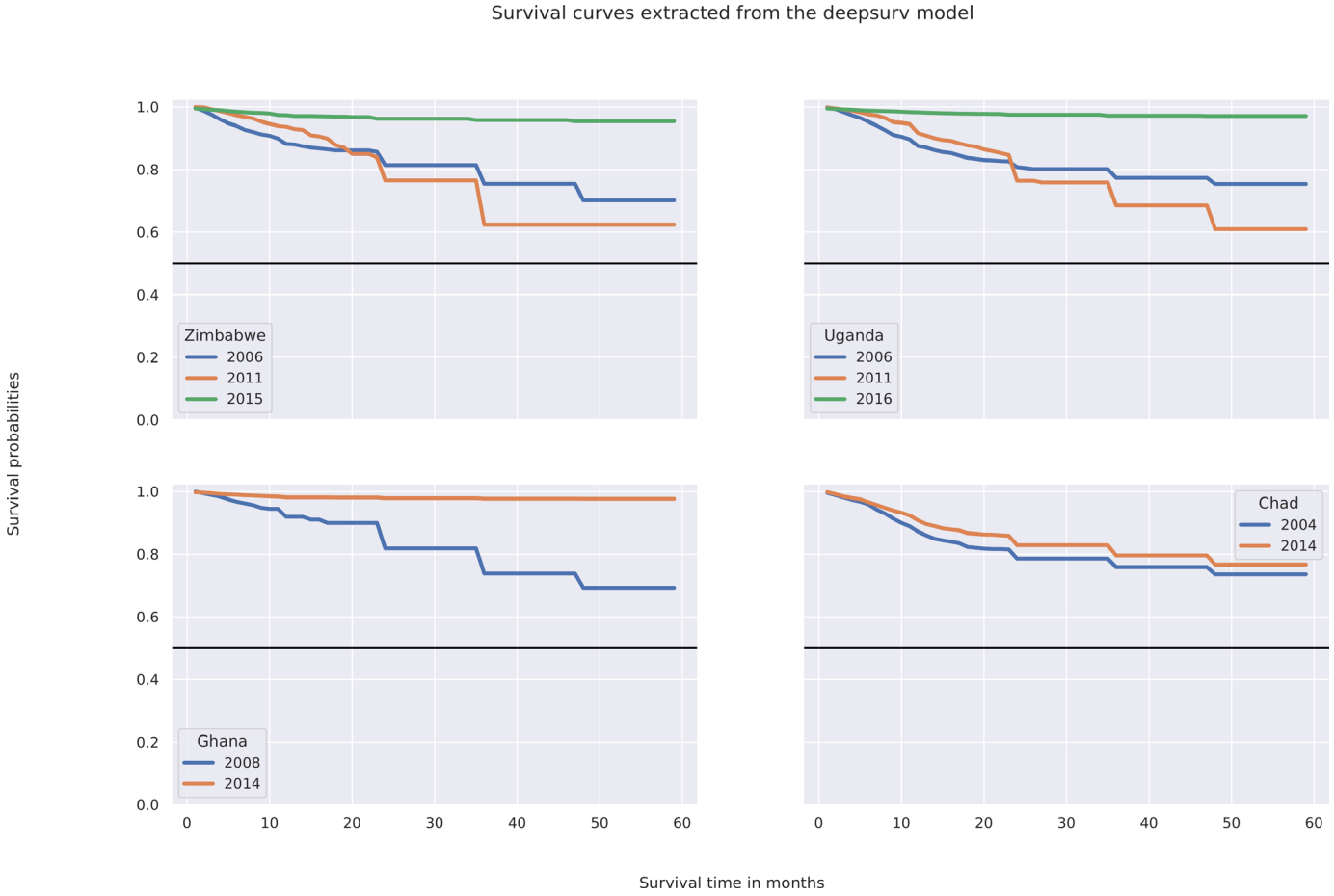

Figure 6 Survival probabilities for the children in the test dataset for Zimbabwe, Uganda, Ghana and Chad obtained from the DeepSurv model.

considered in this study. The survival curves show an improvement in the survival probabilities associated with the four socioeconomic factors for children under the age of 5 in the countries over time. Zimbabwe, in the southern African subregion, had a survival curve for the year 2015 above the survival curves of 2006 and 2011. Uganda, in the East African region, had a survival curve for the year 2001 that is below the survival curve for the year 2016. Ghana, in the West African subregion, had a survival curve for the children under the age of 5 in the year 2014 above that of the year 2008. And lastly, for Chad, in the central subregion, the survival curve for the year 2014 is above that of 2004.

This indicates that there is improvement in the survival outcome associated with the four socioeconomic factors in these countries' over time, especially after 5 or more years after the launch of the MDG.

The countries considered for analysis in the different subregions had a median survival time associated to the four socioeconomic factors for the children in the test dataset of above 5 years; however, we noticed that this improvement has been gradual. For example, a country like Uganda from the East African subregion had a survival curve for the year 2006 that is below the survival curve for the year 2011. It is also shows that the survival curve of the year 2011 is below that of the year 2016.

In Zimbabwe, for the year 2011, the survival curve for the children under the age of 1 year is above that of the children below the same age in 2006. However, the survival curve for children above 1 year in 2011 compared with those above 1 year of age in 2006 are the same. This is expected for short period (2006–2011), however, when we compare the effects of the four factors over a longer period (2006–2015) we can clearly see the distinction between the survival outcomes associated with the four socioeconomic factors over time.

This indicates that there is improvement in the survival outcome associated to the four socioeconomic factors in this country over time. The improvements in the survival outcome associated to these factors over time as evidenced from the results are occurring after the year 2000 where many interventions were implemented to achieve the MDGs, an indicator that these interventions had a positive impact on reducing U5MR.

Lastly, we compared the DeepSurv and RSF models using cross-validated C-indicies to determine which of the two models has a higher predictive performance on the datasets used in this study. These results are, therefore, summarised in figure 7.

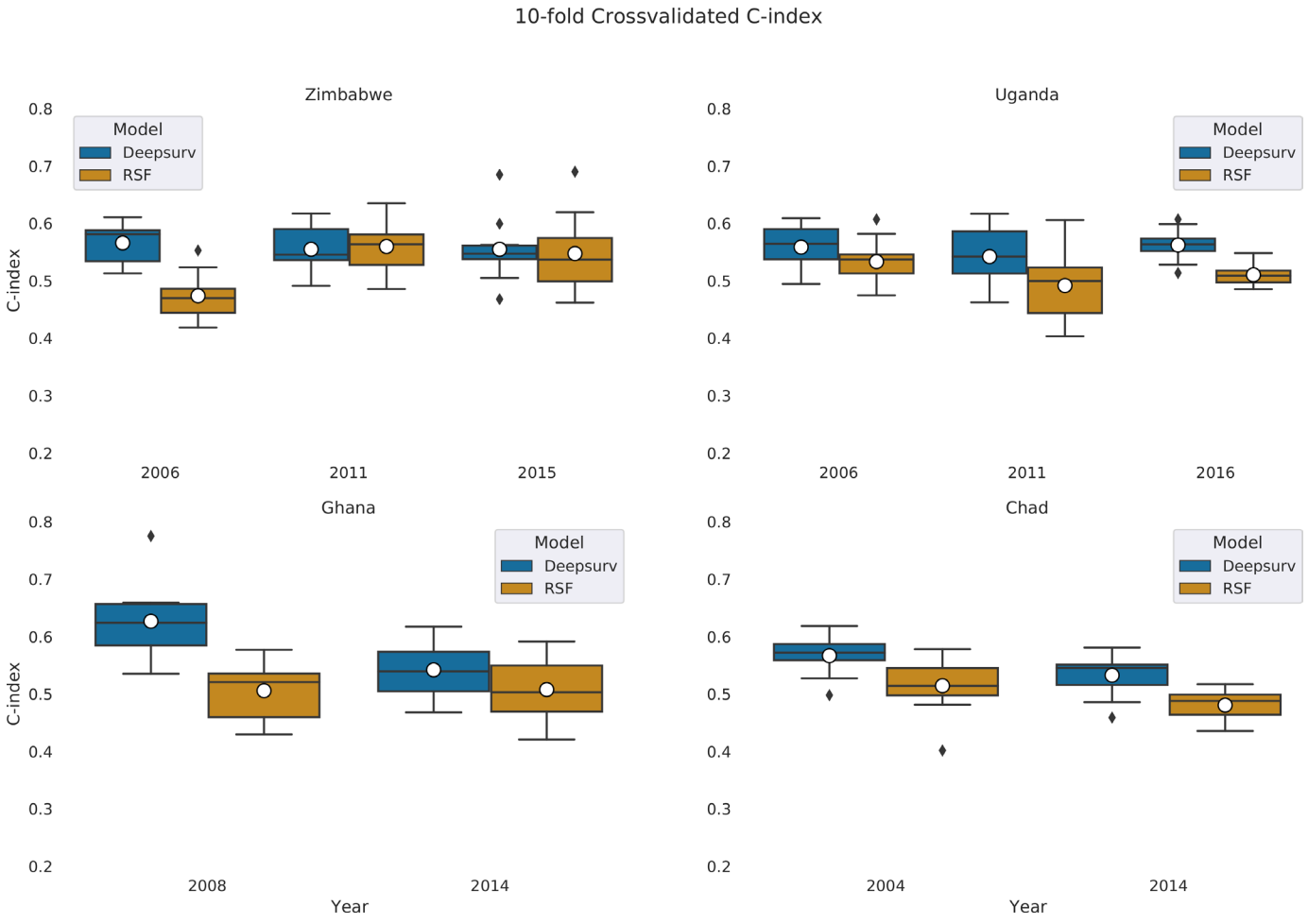

Figure 7 Comparison of predictive performance of the deep survival neural network and the random survival forest (RSF) models on all the datasets considered in this study.

Figure 7 shows that the mean values of the cross-validated C-indices from the deep learning model on all datasets are above the 50% mark, which is an indicator that the model has higher predictive quality compared with the RSF model.

The performance of this model on datasets of a country from each subregion has no clear trend, but what is obvious is that these four socioeconomic factors are still predictive in determining U5MR in sub-Saharan Africa. In fact, in some of the datasets, the model shows a high predictive performance in the recent years. This is an indication that the factors considered in this model are still predictive and associated with U5MR. Therefore, public health policies needed to achieve SDG3 must be designed to target existing inequalities in U5MR caused by these four social economic factors.

DISCUSSION

The study reveals that among the four socioeconomic factors, wealth index (household wealth) and mother's education level are the top contributors of mortality in the countries' datasets considered in this study. Wealth index ranked first in some of the datasets like Zimbabwe (2006), Uganda (2011) and Ghana (2003). It also ranked second in datasets like Zimbabwe (2011 and 2015), Uganda (2006 and 2016), Chad (2008 and 2014) and Ghana (2008 and 2014). Mother's education level was also ranked first in some of the datasets over the period considered, these include Zimbabwe (2015), Uganda (2006 and 2016), and Ghana (2008 and 2014). Place of residence ranked first in datasets like Chad (2004 and 2014).

With a mean C-index value of above 0.5, the deep survival model was the best performing model in predicting U5MR in all the datasets analysed in the study. This implies that the socioeconomic factors included in the model are still very predictive in determining U5MR. Survival curves of the survival outcome associated with the four socioeconomic factors were extracted from the best performing model. These curves are extracted from the deep survival model run on the test dataset, a 20% partition of each of the datasets in the study. For a country like Zimbabwe selected from the Southern African subregion, the recent year, 2015, had survival curves (favourable survival outcome) that were above the survival curves of the earlier years (2006, 2011) on the test data. The general trend in this analysis was that there was a favourable survival outcome associated to the four social economic factors in the recent years compared with the earlier years in the four countries selected from the different subregions.

The main strength of this study is that we used machine learning methods which, when compared with classical statistical models, are very flexible and have fewer assumptions. They are, therefore, adapted to fitting very large datasets with complex relations between predictors and a given response. Another strength of the study is that we are tracking the influence of socioeconomic factors in determining U5MR over time, which has potential to explain how effective our interventions have been. However, the methods used in this study are criticised for being a black box. They may not give an effect size of the factors, and therefore, it is difficult to tell by how much the factor affects the outcome. Another limitation of the study is that the survey data does not include information for mothers who died before the survey, which creates respondent bias.

Our results on the most influential factors associated with U5MR agree with other studies.[2 3 25 52–54] Ezeh *et al*,[54] found that mother's education level and household wealth influenced child survival in Nigeria. A similar study by Adegbosin *et al*,[25] that used deep learning techniques in predicting U5MR in low-inome and middle-income countries, ranked mother's education and household wealth index among the most critical predictors of U5MR. The same study found that deep learning techniques are superior in predicting child survival, and a similar conclusion has been arrived at in other similar studies.[55 56] The only difference in our study is that we were able to extract the survival outcome from the best performing model for each of the countries over time, and presented how the survival outcome associated to the economic factors has improved over time.

In general, there has been a downward trend for U5MR worldwide.[2 54 57 58] Most studies assert that this trend has not occurred evenly in some of the regions. Sub-Saharan Africa is one of those regions with inequalities across countries and social groups. These inequalities in U5MR have evolved over the past 25 years and therefore policymakers must resort to evidence-based policy implementations to achieve the SDG3 target. This study has revealed that machine learning techniques are effective in providing us with such evidence. This study focused on four socioeconomic factors. Among these factors, wealth index and mother's education, were ranked as the most influential in predicting U5MR in the countries used in this study over time. Therefore, policies to achieve SDG3 should directly impact household incomes and girl child education. It is important to note that this study was limited to tracking the ranks of importance of four social economic factors over time and it would be significant to see the changes in the ranks of importance when all the other factors associated with U5MR are included in the study. It would also be vital to see how the survival outcome is improving over time after considering all the other factors that determine U5MR in the region. The study excluded some of the datasets within the countries chosen for analysis, mostly those collected before the year 2000. Including these datasets would lead to us clearly assessing the impact of the interventions that were launched to achieve the MDG to improve the survival outcome of children under the age of 5 in the region.

CONCLUSION

Sub-Saharan Africa has, over the years, implemented policies especially in public health with little or no research to find out which policies would be efficient. This has led to governments and international organisations that are funding these implementations losing much needed resources on inefficient policies. Now, with the availability of datasets like those from the DHSs and the use of machine learning techniques, we can uncover a lot of policy signals. If used well, this information can guide policy-makers on what policies to implement and what sectors to target to achieve the SDG. For example, our study looked at how ranks of importance, the survival outcome, and the predictive nature of four socioeconomic determinants of U5MR have evolved using two machine learning techniques. The results uncovered interesting results that can be used to inform policy on what sectors to target to achieve SDG3. The study revealed that most policies should target reducing poverty levels and aim at increasing literacy levels of the girl child in the regions. The study revealed that past interventions aimed at targeting these four social economic factors are starting to pay-off. This is because, over time, the survival outcome associated with these factors has become more and more favourable.

The DeepSurv model has a higher predictive performance with mean C-index values (between 67% and 80%), above 50%, indicating that these factors are still highly associated with U5MR. Therefore, this study advocates for reviews of the success of these policies using machine learning methods to know where to put the most effort in the implementation process of these programmes targeting some of these factors. The results also show that the deep survival neural network model has a better predictive performance between the two machine learning models.

Twitter Justine B Nasejje @NasejjeJustine

Acknowledgements The first and second authors acknowledges support from the University of the Witwatersrand. The third author acknowledges financial support from the Sub-Saharan Africa Consortium for Advanced Biostatistics training (SSACAB) grant as part of the DELTAS Africa Initiative. The authors acknowledge the DHS Programme for making data available for the countries considered in this study. The authors also acknowledge all the women who participated in the survey together with the teams that conducted the surveys.

Contributors JBN and HM conceptualised the study, JBN conducted the data extraction, JBN and RM trained the models on the datasets and wrote the first draft of the manuscript. HM edited and proofread the document. All authors discussed the results and contributed to the final manuscript.

Funding Grant information: This work was supported through a sub-Saharan Africa Consortium for Advanced Biostatistics training (SSACAB) grant as part of the DELTAS Africa Initiative (107754/Z/15/Z). The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS)'s Alliance for Accelerating Excellence in Science in Africa (AESA), and supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency), with funding from the Welcome Trust (107754/Z/15/Z) and the UK government.

Disclaimer The views expressed in this publication are those of the authors and not necessarily those of AAS, NEPAD Agency, Welcome Trust, or the UK government. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval Permission to use the datasets from all the countries included in the study was granted by the Measure Demographic Health Survey. Ethics approval exemption was granted for the use of these secondary datasets by the University of the Witwatersrand Human Research Ethics Committee (Non-Medical).

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available on reasonable request. Data may be obtained from a third party and are not publicly available. All the datasets used in this study are held by the Demographic and Health Survey program (DHS) and some of the countries' datasets are available on request from the Demographic and Health Survey programme.

ORCID iD

Justine B Nasejje http://orcid.org/0000-0001-8785-8808

REFERENCES

1 Nasejje JB, Mwambi HG, Achia TNO. 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* 2015;15:1003.
2 Tabutin D, Masquelier B, Grieve M. Mortality inequalities and trends in low- and middle-income countries, 1990–2015. *Population* 2017;72:221–95.
3 Van Malderen C, Amouzou A, Barros AJD, *et al*. Socioeconomic factors contributing to under-five mortality in sub-Saharan Africa: a decomposition analysis. *BMC Public Health* 2019;19:760.
4 Mosley WH, Chen LC. An analytical framework for the study of child survival in developing countries. *Popul Dev Rev* 1984;10:25–45.
5 Satagopan JM, Ben-Porat L, Berwick M, *et al*. A note on competing risks in survival data analysis. *Br J Cancer* 2004;91:1229–35.
6 Yohannes T, Laelago T, Ayele M, *et al*. Mortality and morbidity trends and predictors of mortality in under-five children with severe acute malnutrition in Hadiya zone, South Ethiopia: a four-year retrospective review of hospital-based records (2012-2015). *BMC Nutr* 2017;3:18.
7 Sahu D, Nair S, Singh L, *et al*. Levels, trends & predictors of infant & child mortality among Scheduled Tribes in rural India. *Indian J Med Res* 2015;141:709.
8 Meshram II, Arlappa N, Balakrishna N, *et al*. Trends in the prevalence of undernutrition, nutrient and food intake and predictors of undernutrition among under five year tribal children in India. *Asia Pac J Clin Nutr* 2012;21:568–76.
9 Akinyemi JO, Bamgboye EA, Ayeni O. New trends in under-five mortality determinants and their effects on child survival in Nigeria: a review of childhood mortality data from 1990-2008. *African Population Studies* 2013;27:25–42.
10 Kanmiki EW, Bawah AA, Agorinya I, *et al*. Socio-Economic and demographic determinants of under-five mortality in rural Northern Ghana. *BMC Int Health Hum Rights* 2014;14:24.
11 Ayele DG, Zewotir TT, Mwambi H. Survival analysis of under-five mortality using COX and frailty models in Ethiopia. *J Health Popul Nutr* 2017;36:25.
12 Kayode GA, Adekanmbi VT, Uthman OA. Risk factors and a predictive model for under-five mortality in Nigeria: evidence from Nigeria demographic and health survey. *BMC Pregnancy Childbirth* 2012;12:10.
13 Morakinyo OM, Fagbamigbe AF. Neonatal, infant and under-five mortalities in Nigeria: an examination of trends and drivers (2003-2013). *PLoS One* 2017;12:e0182990.
14 Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* 1994;81:515–26.

15 Nasejje JB, Mwambi H, Dheda K, *et al*. A comparison of the conditional inference survival forest model to random survival forests based on a simulation study as well as on two applications with time-to-event data. *BMC Med Res Methodol* 2017;17:115.
16 Faraggi D, Simon R. A neural network model for survival data. *Stat Med* 1995;14:73–82.
17 Ishwaran H, Kogalur UB, Blackstone EH, *et al*. Random survival forests. *Ann Appl Stat* 2008;2:841–60.
18 Yousefi S, Amrollahi F, Amgad M, *et al*. Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. *Sci Rep* 2017;7:11707.
19 LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;521:436–44.
20 Luck M, Sylvain T, Cardinal H. Deep learning for patient-specific kidney graft survival analysis. *arXiv:170510245 [csstat]* 2017.
21 Katzman JL, Shaham U, Cloninger A, *et al*. DeepSurv: personalized treatment recommender system using a COX proportional hazards deep neural network. *BMC Med Res Methodol* 2018;18:24.
22 Sargent DJ. Comparison of artificial neural networks with other statistical approaches: results from medical data sets. *Cancer* 2001;91:1636–42.
23 Xiang A, Lapuerta P, Ryutov A, *et al*. Comparison of the performance of neural network methods and COX regression for censored survival data. *Comput Stat Data Anal* 2000;34:243–57.
24 Mariani L, Coradini D, Biganzoli E, *et al*. Prognostic factors for metachronous contralateral breast cancer: a comparison of the linear COX regression model and its artificial neural network extension. *Breast Cancer Res Treat* 1997;44:167–78.
25 Adegbosin AE, Stantic B, Sun J. Efficacy of deep learning methods for predicting under-five mortality in 34 low-income and middle-income countries. *BMJ Open* 2020;10:e034524.
26 Kumar S, Kumar N, Vivekadhish S. Millennium development goals (MDGs) to sustainable development goals (SDGs): addressing unfinished agenda and strengthening sustainable development and partnership. *Indian J Community Med* 2016;41:1.
27 Nasejje JB, Mwambi H. 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 Res Notes* 2017;10:459.
28 Breiman L, Friedman J, Stone CJ. *Classification and regression trees*, 1984.
29 Morgan JN, Sonquist JA. Problems in the analysis of survey data, and a proposal. *J Am Stat Assoc* 1963;58:415–34.
30 Gordon L, Olshen RA. Tree-structured survival analysis. *Cancer Treat Rep* 1985;69:1065–9.
31 Bou-Hamad I, Larocque D, Ben-Ameur H. A review of survival trees. *Stat Surv* 2011;5:44–71.
32 Breiman L. Random forests. *Mach Learn* 2001;45:5–32.
33 Dietterich TG. *Ensemble learning. The Handbook of brain theory and neural networks*. Arbib MA, 2002.
34 Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: a conditional inference framework. *Journal of Computational and Graphical Statistics* 2006;15:651–74.
35 Wright MN, Dankowski T, Ziegler A. Unbiased split variable selection for random survival forests using maximally selected RANK statistics. *Stat Med* 2017;36:1272–84.
36 Wright MN, Ziegler A. ranger : A Fast Implementation of Random Forests for High Dimensional Data in *C++* and *R*. *J Stat Softw* 2017;77.
37 R Core Team. R: a language and environment for statistical computing, 2013. Available: https://www.R-project.org/
38 Ishwaran H, Kogalur UB, Kogalur MU. *Package 'randomSurvivalForest'*, 2013.
39 Cox DR. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B* 1972;34:187–202.
40 Fotso S. *PySurvival: open-source package for survival analysis modeling.*", 2019.
41 Harrell FE, Lee KL, Califf RM, *et al*. Regression modelling strategies for improved prognostic prediction. *Stat Med* 1984;3:143–52.
42 Gönen M, Heller G. Concordance probability and discriminatory power in proportional hazards regression. *Biometrika* 2005;92:965–70.
43 Pencina MJ, D'Agostino RB. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Stat Med* 2004;23:2109–23.
44 Santos MY, Oliveira e Sá J, Andrade C, *et al*. A big data system supporting Bosch Braga industry 4.0 strategy. *Int J Inf Manage* 2017;37:750–60.
45 Schwarz DF, König IR, Ziegler A. On safari to random jungle: a fast implementation of random forests for high-dimensional data. *Bioinformatics* 2010;26:1752–8.
46 Jones Z, Linder F. Exploratory data analysis using random forests. *Prepared for the 73rd annual MPSA conference*, 2015.
47 Ishwaran H. Variable importance in binary regression trees and forests. *Electron J Stat* 2007;1:519–37.
48 Ishwaran H, Kogalur UB, Gorodeski EZ, *et al*. High-Dimensional variable selection for survival data. *J Am Stat Assoc* 2010;105:205–17.
49 Strobl C, Boulesteix A-L, Zeileis A, *et al*. Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics* 2007;8:25.
50 Wright MN, Ziegler A, König IR. Do little interactions get lost in dark random forests? *BMC Bioinformatics* 2016;17:145.
51 Strobl C, Boulesteix A-L, Kneib T, *et al*. Conditional variable importance for random forests. *BMC Bioinformatics* 2008;9:307.
52 Rutstein S, Winter R, Staveteig S. Urban child poverty, health, and survival in Low-and middle-income countries. *PAA 2017 Annual Meeting*, 2017.
53 Kunst AE, Mackenbach JP. The size of mortality differences associated with educational level in nine industrialized countries. *Am J Public Health* 1994;84:932–7.
54 Ezeh OK, Agho KE, Dibley MJ, *et al*. Risk factors for postneonatal, infant, child and under-5 mortality in Nigeria: a pooled cross-sectional analysis. *BMJ Open* 2015;5:e006779.
55 Taylor RA, Pare JR, Venkatesh AK, *et al*. Prediction of in-hospital mortality in emergency department patients with sepsis: a local big data-driven, machine learning approach. *Acad Emerg Med* 2016;23:269–78.
56 Panesar SS, D'Souza RN, Yeh F-C, *et al*. Machine learning versus logistic regression methods for 2-year mortality prognostication in a small, heterogeneous glioma database. *World Neurosurg* 2019;2:100012.
57 Kimani-Murage EW, Fotso JC, Egondi T, *et al*. Trends in childhood mortality in Kenya: the urban advantage has seemingly been wiped out. *Health Place* 2014;29:95–103.
58 Sousa A, Hill K, Dal Poz MR. Sub-national assessment of inequality trends in neonatal and child mortality in Brazil. *Int J Equity Health* 2010;9:21.