BMJ Open

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

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2021-049786
Article Type:	Original research
Date Submitted by the Author:	02-Feb-2021
Complete List of Authors:	Nasejje, Justine B; University of the Witwatersrand, Statistics and Actuarial Science Mbuvha, Rendani; University of the Witwatersrand, Statistics and Actuarial Science Mwambi, Henry; University of Kwazulu-Natal, School of Mathematics, Statistics and Computer Science
Keywords:	Health policy < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, PUBLIC HEALTH, Community child health < PAEDIATRICS

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

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

1 School of Statistics and Actuarial science, University of Witwatersrand, Jan Smuts Avenue, Johannesburg, Gauteng, South Africa

2 School of Statistics, Mathematics and Computer Science, University of KwaZulu-Natal, King Edward Avenue, Pietermaritzburg, South Africa

* Corresponding author E-mail: justine.nasejje@wits.ac.za

Abstract

Objectives Use machine learning algorithms to track how the ranks of importance and predictive nature of four socioeconomic determinants of U5MR in sub-Saharan Africa (place of residence, mother's level of education, wealth index, and sex of the child) have evolved overtime.

Settings It is a Cross-section study, and we analyzed data from Demographic Health Surveys (DHS)

Participants Data were drawn from 16 sub-Saharan countries, four countries selected from each sub-region. A total of n= 521,873 children were drawn

Interventions Reducing U5MR was the fourth Millennium Development Goals (MDGs) drafted in the year 2000, and the world sprung into action to achieve it and now it appears within the third Sustainable Development Goal (SDG3)

Primary and secondary outcomes The primary outcome variable is U5MR; secondary outcomes are rank importance of the socioeconomic factors over-time and comparing the two machine learning models; random survival forest (RSF) and the deep survival neural network (DeepSurv) in predicting U5MR

Results Wealth index ranks top among the factors in majority of the countries in the region, followed by mother's education level. Sex of the child is found to have declining importance. The DeepSurv has a higher predictive performance with mean concordance indexes above 50% compared to the RSF model. Generally, the four factors show favorable U5MR over-time. Hence, affirming that past interventions aimed at targeting these factors in the region are beginning to payoff.

Conclusions The study has revealed that policies aimed at reducing poverty levels and increasing literacy levels of the girl child in the region should still be favoured. Policies on closing the gender gap are starting to pay-off. It also shows that deep learning models are efficient in predicting U5MR and should therefore be used in this big data era to draft evidence based policies aimed at achieving SDG3 in the region.

Strengths and limitations of the study

(1) The main strength of this study is that, machine learning methods compared to classical statistical models are very flexible, that is to say, have fewer assumption and are therefore adopted to fitting very large datasets with complex relations between predictors and a given response or outcome.

(2) A limitation of this study is that, machine learning models do not give an effect size of the factors and therefore it is very difficult to tell by how much the factor affects the outcome.

Introduction

The probability of a child dying before the age of 5 years (U5MR) is a global indicator of societal and national development as it serves as a key marker of health equity and access [2]. The fourth Millennium Development Goal (MDG 4) which previously stated that, reducing under-five mortality by two-thirds in the period between 1990 and 2015 now appears in the third Sustainable Development Goal (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 still exists substantial inequalities between subgroups of the population within countries [22,23]. These sub-groups are based on factors such as, wealth index, maternal factors such as education level, place of residence, sex of the child, among others. The Mosley and Chen framework [24], categorizes these social economic factors as the distal determinants of child mortality [24].

Classical statistical parametric regression models such as the logistic regression model, semi-parametric models like the Cox proportional hazard models (CPH) and generalized additive models have been widely used to study determinants of U5MR [1–8]. A study by [4] 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 socio-demographic and economic factors were observed over the time period. In a study by [5], it was concluded that mother's education level and sex of the child were among the factors responsible for trends and differentials of U5MR in Ethiopia. Similar studies in Nigeria concluded that place of residence (rural or urban) was an important risk factor in determining U5MR [9]. Mothers' education, place of residence and sex of the child were also found significant in influencing U5MR trends in Nigeria [10]. Although the CPH and the logistic regression models are very robust, they are often criticised for their restrictive assumptions and hence may lead to bias if care is not taken when preparing the data for analysis [11]. Classical machine learning approaches which include: nearest neighbours, neural networks, kernel methods, penalized least squares and data partitioning methods such as decision trees (CART) and Random forests are among the alternative approaches to parametric and semi-parametric classical models [12–14]. Recently, deep learning methods which are advances in neural networks have been recommended for analysing survival data [15–21]. These machine learning models are known to be very flexible compared to the statistical models like the CPH model [18–21,56]. A recent study by [56] recommended the use deep learning models in understanding the determinants of U5MR in low and middle income countries.

This study uses two machine learning models; the random survival forest model to track how ranks of importance of four socioeconomic factors in determining U5MR have evolved and the deep survival model aimed at identifying how predictive these socioeconomic factors are in determining U5MR in sub-Saharan Africa over the time period considered. These factors include; place of residence, mother's level of education, wealth index, and sex of the child.

Studying how the rank in importance of these factors in determining U5MR has evolved over time can help redirect resources to the right sectors and hence be on-course to achieving SDG3. In this study, therefore we fit a random survival forest and deep survival neural network model to understand how the rank of importance and predictive nature of these socioeconomic factors in determining U5MR in sub-Saharan Africa has evolved over time. The random survival forest 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 drivers of U5MR in this region by looking at the changes in the survival outcome associated to these factors over-time.

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

Methods

Data

Datasets of completed Standard Demographic and Health Surveys (DHS) from four countries in each of the four regions (Southern, Central, Eastern and Western Africa) in sub-Saharan Africa are used. DHS funded by USAID, UNFPA, UNICEF, Irish Aid and the United Kingdom government have over the years (since 1988), provided datasets which are 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 [56]. More information about the sampling design, data collection and processing details are described on the DHS program website. The outcome variable is under-five survival time and this information was obtained from the birth history of interviewed women aged between 15 to 49 years of age. All the datasets used in this analysis comprised of children dead or alive, born in the period of five years preceding the date of the survey. This is done 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 education, wealth index of the household, and sex of the child. The study considered four countries from each region as shown in Table 1 below. The datsets considered for the analysis are summarised in Table 2. The total number of children in some of the datasets and the number of deaths are also summarised in Table 2.

Table 1. The standard Demographic and Health Survey datasets used for this study by region

Southern region				Eastern Region			
Zimbabwe	**Malawi**	**Namibia**	**Zambia**	**Uganda**	**Kenya**	**Tanzania**	**Ethiopia**
1999	2000	1992	1996	2001	1998	1999	2000
2006	2004	2000	2001	2006	2003	2004	2005
2011	2010	2006	2007	2011	2008	2010	2011
2015	2015	2013	2013	2016	2014	2015	2016
Western region				**Central region**			
Senegal	**Ghana**	**Benin**	**Mali**	**Cameroon**	**DRC**	**Gabon**	**Chad**
1992	1998	2001	1995	1991	2007	2000	1996
1997	2003	2006	2001	1998	2013	2012	2004
2005	2008	2011	2006	2004			2014
2010	2014	2017	2012	2011			

Datasets available for each of the four selected countries in the four regions of sub-Saharan Africa and the year the survey was conducted.

Table 2. The total number of deaths under the age of five in each dataset (Supplementary material)

Patient and Public Involvement

There are no patients involved in this study

Exploratory data analysis

Figures 1 and 2 show sample exploratory data analysis plots for Zimbabwe and Ghana stratified by the socioeconomic factors considered in our study.

Fig 1. EDA plots for the covariates. Bars indicate number of children under five years for each covariate. Colors correspond to class membership within each covariate

Fig 2. EDA plots for the covariates. Bars indicate number of children under five years for each covariate. Colors correspond to class membership within each covariate

Models

The CPH model is the most frequently used model to analyse survival data [1,2]. 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 also involve interactions between 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 random survival forests which can enable someone detect complex relationships in survival datasets have been employed in recent years [12]. These methods have achieved high accuracy in predicting the survival outcomes when applied to survival datasets to identify factors affecting U5MR [25]. Despite the fact that they have exhibited a good performance in predicting survival outcomes, there are few studies aimed at understanding factors associated to U5MR that have embraced these methods [12,25]. Recently, with the advancement of the machine learning methods, deep learning methods have also been added to the tool box of methods to analyse survival data [18]. The fact that most datasets collected have complex structures, using models that have very strict assumptions may lead to bias and hence misleading policy implementations. In this study, we apply two machine learning models on datasets from sub-Saharan Africa aimed at understanding how the rank of importance and predictive nature of the socioeconomic factors in determining U5MR has evolved over time. These two models are; the random survival forest [14] and the deep survival neural network model (DeepSurv) [18]. It is important to note that a deep survival neural network model is being used to identifying how the predictive nature of the four socioeconomic determinants of U5MR has evolved over time by looking at the changes in the survival outcome (Under five survival times) associated to these four factors over the time period considered for each country.

Random survival forests

Random survival forests are an extension of regression trees formally presented by [26] to survival data. These methods have been found to be the most desirable methods in addressing the above mentioned challenges of the CPH model. The algorithm of the random survival forest model by [26] and that of the conditional inference survival forest model are described in detail below but first, we describe the survival tree algorithm an important building block of the forest.

Survival trees

The regression tree algorithm for right censored data is an extension of the CART algorithm by [26,51,53,54]. Below is the general algorithm for survival trees [53,54]

Algorithm1 : 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 ($\alpha$ and $\beta$) 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 i.e., has no less than $d_0 > 0$ unique observed events.
```

An RSF model is a collection of survival trees because a single tree is always not a good probability estimator due to its short comings of giving unstable estimators [55,57]. Researchers have over the years recommended the growing of an entire forest as the solution to the shortcomings of a single tree. The algorithm for building an RSF model as presented by [58] is given below as follows

Algorithm2 : 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.
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 maximizes 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 deaths.
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.
```

Note that the node size is restricted such that the number of unique events at a node does not drop below the minimum number. The Random survival model was fit in the R-software [60,61] with each forest consisting of 200 trees.

Neural network survival models

Non-linear models like artificial neural networks are increasingly becoming popular as additional models to the tool box of models aimed at predicting survival outcomes. They look very promising especially in application to large datasets that could be having a large number of covariates with non-linear effects on the survival outcome. It is important to note that neural networks are only very good for predicting outcomes but not able to give explanations or quantify covariate effects on the outcomes. Initially, a single hidden layer feed-forward neural network were fitted to survival data and their performance in predicting survival outcomes provided mixed results [18–21]. 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 [15–17]. Instead of a 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 [32]. Therefore before describing the neural network, we give a gentle introduction to the CPH model.

Cox proportional hazards model

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

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

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 $\hat{h}_\beta(X) = \beta' X$. The estimates $(\hat{\beta})$ of the parameters (β) are obtained by maximising the partial likelihood. Suppose that there are k distinct event times, and $t_1 < t_2 < .... < t_k$ represent the ordered distinct event times, the partial likelihood is given as

$$L(\beta) = \prod_{i=1}^{k} \frac{\exp\left(\hat{h}_\beta (X_i)\right)}{\sum_{j \in \Re(t_i)} \exp\left(\hat{h}_\beta (X_i)\right)}. \quad (2)$$

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 ReLU is recommended in modern neural networks as the transformation or activation function to compute hidden layer values. This is defined as

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

In this study, however, the Scaled Exponential Linear Unit (SELU) is used as an activation function because of its advantages over the ReLU. ReLUs 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 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 zero. If this happens to many cells in your network, the power of the trained network stays below its theoretical capabilities. It is given as

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

Where $\gamma > 0$ and $\lambda > 0$ 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 $\hat{h}_\beta(X) = \beta^0 X$ in equation 2 by the output of $\hat{h}_\theta(X) = \exp(g(X, \theta))$ of the neural network. The proportional hazards model becomes;

$$h_\theta(X_i) = \exp(g(X_i, \theta)). \quad (4)$$

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

$$L(\theta) = \prod_{i:\delta_i=1} \frac{\exp\left(\hat{h}_\theta (X_i)\right)}{\sum_{j \in \Re(t_i)} \exp\left(\hat{h}_\theta (X_i)\right)}. \quad (5)$$

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

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

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

$$l\left(\theta\right) = -\frac{1}{n_{\delta_1}} \sum_{i:\delta_i=1} \left(\hat{h}_\theta\left(X_i\right) - log \sum_{j\in\Re(t_i)} \exp\left(\hat{h}_\theta\left(X_j\right)\right)\right) + \alpha \left\|\theta\right\|_2^2 \qquad (7)$$

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 regularization parameter for the L_2 norm. Gradient descent optimization is used to find the weights of the network which minimise the loss function. The DeepSurv neural network architecture adapted for this study is described in detail by [18]. The figure below shows its architecture. It is a deep feed-forward neural network implemented as

Fig 3. DeepSurv architecture [18]

DeepSurv in *Theano* with the Python package *Lasagne* by [18]. For our study, observed social economic 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 dropout 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 $\exp(-8)$ are used.

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 [33,34]. Note that censored observation cannot be compared with any observed event time because it's exact event time is unknown; however, any other pair of observations are called comparable [35]. If predicted survival outcomes are denoted by $\hat{Y}$, the C-index is given by

$$C = \frac{\sum_{i:\delta_i=1} \sum_{y_i<y_j} I\left(\hat{Y}_i < \hat{Y}_j\right)}{\text{Number of Comparable Pairs}} \qquad (8)$$

In survival analysis, shorter survival time means smaller predicted outcome. C-index value of above 0.5 means better agreement among comparable pairs.

Over-fitting 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 mostly commonly used methods to evaluate the predictive performance of machine learning models. The test error is known to be very 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 also since the test data is a subset of the train set, this independence is not guaranteed. This makes this method unreliable. Hence $K - fold$ crossvalidation is recommended. $K - fold$ crossvalidation divides the data into K folds and ensures that each fold is used as a testing set at some point [36]. In this study, we use 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 combined 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 fold of the 10 folds have been used as the testing set.

Measures of covariate importance

To understand which factors are important in influencing predictions, the random survival forests model has a measure of estimating importance of each covariate. It is generally referred to as the variable importance measure (VIMP) [37–40]. Variables are selected on the basis of their importance in predicting the survival outcome. The basic measure of variable importance is by counting the number of times the predictor is selected by each tree in the whole forest [41]. Different measures of variable importance exist in literature and have been implemented in the random forest algorithms [26,41–43]. 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. For illustration, first consider a tree grown and its prediction accuracy ($\hat{e}$), calculated using the out-of-bag (OOB) observations. Secondly, randomly permute the values of the factor of interest, (X_j) for all individuals. Note that permutation breaks the original relationship of the covariate with the survival outcome. Obtain a new value for prediction accuracy, ($\hat{e}_j$) using OOB observations. Compare $\hat{e}_j$, with $\hat{e}$ of the original classification for covariate, X_j. Calculate, argmax$\{0; \hat{e}_j - \hat{e}\}$. The difference between the accuracy before and after permutation provides the importance of the covariate, X_j 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 [42–44].

Results

Fig 4. Ranks of importance for the four social economic factors in predicting U5MR in Malawi over the period of 11 years

Fig 5. Ranks of importance for the four social economic factors in predicting U5MR in Namibia over the period of 7 years

Fig 6. Ranks of importance for the four social economic factors in predicting U5MR in Zimbabwe over the period of 9 years

Fig 7. Ranks of importance for the four social economic factors in predicting U5MR in Kenya over the period of 11 years

Fig 8. Ranks of importance for the four social economic factors in predicting U5MR in Ethiopia over the period of 11 years

Fig 9. Ranks of importance for the four social economic factors in predicting U5MR in Senegal over the period of 5 years

Fig 10. Ranks of importance for the four social economic factors in predicting U5MR in Tanzania over the period of 11 years

Fig 11. Ranks of importance for the four social economic factors in predicting U5MR in Ghana over the period of 6 years

Fig 12. Ranks of importance for the four social economic factors in predicting U5MR in Benin over the period of 11 years

Fig 13. Ranks of importance for the four social economic factors in predicting U5MR in Mali over the period of 6 years

Fig 14. Ranks of importance for the four social economic factors in predicting U5MR in Cameroon over the period of 7 years

Fig 15. Ranks of importance for the four social economic factors in predicting U5MR in DRC over the period of 6 years

Figures 4-19 show that wealth index has been ranked as the top predictor of under-five survival in most of the countries considered in this study over the period of atleast 5

Fig 16. Ranks of importance for the four social economic factors in predicting U5MR in Gabon

Fig 17. Ranks of importance for the four social economic factors in predicting U5MR in Uganda over the period of 10 years

Fig 18. Ranks of importance for the four social economic factors in predicting U5MR in Zambia over the period of 6 years

Fig 19. Ranks of importance for the four social economic factors in predicting U5MR in Chad over the period of 10 years

years. This result is in agreement with a study by [45] which studied the changes in socioeconomic inequalities in low and middle income countries in the 2000s. It is also clear from our results for some of the countries that the sex of the child is ranking last over time. The other factor that is considered as important is the mother's education level.

Fig 20. Survival probabilities for the children in the test dataset on the Southern Africa datasets obtained from the deepsurv model

Fig 21. Survival probabilities for some of the children on the Eastern Africa datasets obtained from the deepsurv model

Fig 22. Survival probabilities for the children in test dataset on the Western Africa datasets obtained from the deepsurv model

Fig 23. Survival probabilities for the children in the test dataset on the Central Africa datasets obtained from the deepsurv model

Figures 20-23 shows survival curves of the survival outcome(under-five survival time) associated to the four socioeconomic factors extracted from the deep learning survival model for the test datasets obtained from the datasets of all the sub-regions considered in this study. The survival curves show an improvement in the survival probabilities associated to the four socioeconomic factors for the children under the age of five in the countries over-time. Most of these countries in the different sub-regions had a median survival time associated to the four socioeconomic factors for the children in the test dataset of above five years; however, we notice that this improvement has been gradual. For example, a country like Uganda in the East African region had a survival curve for the year 2001 that is below the survival curve for the year 2016. This is an indicator that there is improvement in the survival outcome associated to the four socioeconomic factors in this country over-time. In some countries within given sub-regions have a median survival time associated to the four socioeconomic factors for the children in the test dataset of below five years on the earlier years of the DHS studies but we later notice the improvement in the median survival time over-time. For example, in Malawi, the median survival time of these children was 37 months in the year 2000 but this improved gradually. In the year 2015, we observe that all the children in the test dataset survived beyond five years of age. A similar phenomenon existed in Mali where the median survival time for the children under the age of five years in the test dataset was about 22 months in 1995 but it later improved to be above five years of age in the year 2012. Most of these improvements in the survival outcome associated to these factors continue after the year 2000 where many interventions were implemented to achieve the MDGs, an indicator that these interventions had a positive impact on U5MR.

Figures 24-27 show that the values of the concordance index 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 to the random survival forest model.

Fig 24. Comparison of predictive performance of the deep survival neural network and the random forest models on Southern Africa datasets

Fig 25. Comparison of predictive performance of the deep survival neural network and the random forest models on Eastern Africa datasets

Fig 26. Comparison of predictive performance of the deep survival neural network and the random forest models on Western Africa datasets

Fig 27. Comparison of predictive performance of the deep survival neural network and the random forest models on Central Africa datasets

The performance of this model on each region has no clear trend but what is obvious is that these four social economic factors are still predictive in determining U5MR in sub-Saharan Africa. Infact in some of the regions the model shows a high predictive performance in the recent years. This is an indication that the factors considered in this model are highly predictive and associated to U5MR and therefore public health policies to achieve SDG3 should be designed to target inequalities based on these factors that exist within each countries in the region.

Discussion

There has been a downward trend for U5MR worldwide [22, 46, 47]. Most studies assert that this trend has not occurred evenly in some of the regions. These inequalities in U5MR have evolved over the past 25 years and therefore policy makers have to resort to evidence based policy implementations to achieve the SDG3 target. Sub-Saharan Africa is one of those regions with inequalities across countries and social groups. This study was aimed at uncovering how the rank of importance and predictive nature of the four socioeconomic factors in determining U5MR have evolved over time in this region. Wealth index (household wealth) and Mother's education level are ranked to be the main contributors of mortality in most of the countries in this study. In-fact in countries like Mali, Kenya, Ethiopia, Senegal, Benin, Gabon, DRC and Mali, wealth index was the main contributor to U5MR over the period considered for each of the country. Mother's education level was also ranked first in some of the datasets over the period considered, these countries include, Cameroon, Ghana, Zimbabwe, Namibia and Uganda. Place of residence ranked first in countries like Zambia and Chad. Our results are in agreement with studies by [22, 23, 45, 48, 56]. Policies to achieve SDG3 should directly impact household incomes and girl child education. The sex of the child consistently ranks last in most of the datasets, this could be an indication of how policies to close the gender gap are starting to pay off [49,50]. With a concordance index value of above 0.5, the deep survival model was predictive in all the datasets used. This implies that the social economic factors included in the model are still very predictive in determining U5MR within the region. Survival curves of the survival outcome associated to the four social economic 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 dataset in the study. For the Southern African sub-region, it is clear that Zimbabwe and Namibia in the recent years 2015 and 2013, respectively, had survival curves (favourable survival outcome) that were above the survival curves of the earlier years (2006, 2011, 2000,2006) on the test data. Countries like Malawi and Zambia had the worst survival outcome on the test data for the years 2000 and 2013, respectively. Malawi had a median survival time of about 28 months in 2000 and Zambia had a median survival time of 46 months in 2013. It is very concerning to see such a trend in Zambia given that 2013 is quite recent but 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 to the earlier years in majority of the countries in the different sub-regions.

Conclusion

Sub-Saharan Africa has over the years blindly 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 implementation lose a lot of resources on inefficient policies. Now with the availability of datasets like those from the Demographic health surveys and the use of machine learning techniques, we can uncover a lot of policy signals. If used well, this information can guide policymakers on what policies to implement and what sectors to target inorder to achieve the sustainable development goals. In our study for example, we have looked at how ranks of importance and the predictive nature of four social economic determinants of U5MR have evolved over time using two machine learning techniques. The results have uncovered interesting results that can be used to inform policy on what sectors to target inorder to achieve SDG3. The study has revealed that most of the policies should target reducing poverty levels and also aim at increasing literacy level of the girl child in the region. The study has also revealed that the past interventions aimed at targeting these four social economic factors are starting to pay-off. This is because over-time the survival outcome associated to these factors has become more and more favourable. That is to say, the survival curves on the test data for the earlier years are below those of the recent years in majority of the countries considered in the study. For example in Mali, the survival curve for the year 1995 is below the survival curve for the year 2001. This is an indication of favourable survival outcome associated to the four factors in Mali over-time. This trend is existent in many other countries in the different sub-regions. This does not imply that policies targeting these factors should stop, the fact that the DeepSurv model has a predictive perfomance of above 50%, these factors are still highly associated to U5MR. This study is therefore advocating for reviewing the success of these policies using machine learning methods to know where to put much effort along the implementation process of these policies targeting some of these factors. The results also show that among the two machine learning model, the deep survival neural network model has a better predictive performance compared to the random survival forest model.

Availability of data

All the datasets used in this study are held by the Demographic and Health Survey program and some of the countries' datasets are available on request from the Demographic and Health Survey program.

Ethics declarations

Ethics approval and consent to participate

The ethical statement for all the datasets used in this study is available on the DHS ethical clearance certificate and it states that: The IRB-approved procedures for DHS public-use datasets do not in any way allow respondents, households, or sample communities to be identified. There are no names of individuals or household addresses in the data files. The geographic identifiers only go down to the regional level (where regions are typically very large geographical areas encompassing several states/provinces). Each enumeration area (Primary Sampling Unit) has a PSU number in the data file, but the PSU numbers do not have any labels to indicate their names or locations. In surveys that collect GIS coordinates in the field, the coordinates are only for the enumeration area (EA) as a whole, and not for individual households, and the measured coordinates are randomly displaced within a large geographic area so that specific enumeration areas cannot be identified.

Consent for publication

The DHS programme collects data according to the rules and guidelines stipulated by WHO World Health Survey on consent from the participants stated below.

Participation in the survey is voluntary and the respondent can refuse to be interviewed. The interviewer is responsible for explaining what the survey is about, providing all the necessary information, and making sure the respondent understands the implications of his/her participation before giving his/her consent. The information given should be simple and clear and adapted to the respondent's level of understanding. Consents must be documented by asking the respondents to sign an Informed Consent Forms (Household Informant Consent Form; Individual Consent Form) before doing the interview. These forms must mention who will be doing the study, the types of questions that will be asked, why the study is being done, and who will have access to the information provided. The interviewer must check that the respondent has read and understood the form before signing, and should offer to go over it with him /her emphasizing the different items mentioned. If the respondent is illiterate or unable to read for himself/herself (e.g. due to a visual impairment), the form will be read and explained to him/her. In cases where it is not appropriate for the respondent to sign the form, the interviewer alone will sign the form. In cases where the respondent is being dissuaded from, or coerced into, participating in the study by a third party such as a spouse, relative or any other member in the community, the interviewer should make it clear that it is the respondent alone who must decide whether or not she/he wishes to be interviewed.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

All authors have read and reviewed the manuscript.

Acknowledgements

The first author acknowledges financial support from the University of the Witwatersrand and that from the TWAS-DFG Cooperation Visits Programme. RM is funded by the Google PhD fellowship. The authors acknowledge the DHS Program 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.

References

1. Satagopan JM, Ben-Porat L, Berwick M, Robson M, Kutler D, Auerbach AD. A note on competing risks in survival data analysis. British Journal of Cancer. 2004;91(7):1229–1235.

2. 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(1):1003.

3. Yohannes T, Laelago T, Ayele M, Tamrat T. 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 Nutrition. 2017;3(1):18.

4. Sahu D, Nair S, Singh L, Gulati B, Pandey A. Levels, trends & predictors of infant & child mortality among Scheduled Tribes in rural India. The Indian journal of medical research. 2015;141(5):709.

5. Meshram II, Arlappa N, Balakrishna N, Rao KM, Laxmaiah A, Brahmam GNV, 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 Pacific journal of clinical nutrition. 2012;21(4):568.

6. 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(1).

7. Kanmiki EW, Bawah AA, Agorinya I, Achana FS, Awoonor-Williams JK, Oduro AR, et al. Socio-economic and demographic determinants of under-five mortality in rural northern Ghana. BMC international health and human rights. 2014;14(1):24.

8. Ayele DG, Zewotir TT, Mwambi H. Survival analysis of under-five mortality using Cox and frailty models in Ethiopia. Journal of Health, Population and Nutrition. 2017;36(1):25.

9. 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 and childbirth. 2012;12(1):10.

10. Morakinyo OM, Fagbamigbe AF. Neonatal, infant and under-five mortalities in Nigeria: An examination of trends and drivers (2003-2013). PloS one. 2017;12(8).

11. Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. Biometrika. 1994;81(3):515–526.

12. Nasejje JB, Mwambi H, Dheda K, Lesosky M. 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 Medical Research Methodology. 2017;17(1):115.

13. Faraggi D, Simon R. A neural network model for survival data. Statistics in Medicine. 1995;14(1):73–82.

14. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. Annals of Applied Statistics. 2008;2(3):841–860.

15. Yousefi S, Amrollahi F, Amgad M, Dong C, Lewis JE, Song C, et al. Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. Scientific Reports. 2017;7(1):11707.

16. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–444.

17. Luck M, Sylvain T, Cardinal H, Lodi A, Bengio Y. Deep Learning for Patient-Specific Kidney Graft Survival Analysis. arXiv:170510245 [cs, stat]. 2017;.

18. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Medical Research Methodology. 2018;18(1):24.

19. Sargent DJ. Comparison of artificial neural networks with other statistical approaches: results from medical data sets. Cancer. 2001;91(8):1636–1642.

20. Xiang A, Lapuerta P, Ryutov A, Buckley J, Azen S. Comparison of the performance of neural network methods and Cox regression for censored survival data. Computational Statistics & Data Analysis. 2000;34(2):243–257.

21. Mariani L, Coradini D, Biganzoli E, Boracchi P, Marubini E, Pilotti S, 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 Research and Treatment. 1997;44(2):167–178.

22. Tabutin D, Masquelier B, Grieve M, Reeve P. Mortality Inequalities and Trends in Low- and Middle-Income Countries, 1990–2015. Population, English edition. 2017;72(2):221 – 295.

23. Van Malderen C, Amouzou A, Barros AJD, Masquelier B, Van Oyen H, Speybroeck N. Socioeconomic factors contributing to under-five mortality in sub-Saharan Africa: a decomposition analysis. BMC Public Health. 2019;19(1):760.

24. Mosley WH, Chen LC. An Analytical Framework for the Study of Child Survival in Developing Countries. Population and Development Review. 1984;10:25–45.

25. 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 research notes. 2017;10(1):459.

26. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees; 1984.

27. Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical statistics. 2006;15(3):651–674.

28. Hothorn T, Hornik K, Zeileis A. ctree: Conditional inference trees. The Comprehensive R Archive Network. 2015; p. 1–34.

29. Wright MN, Dankowski T, Ziegler A. Unbiased split variable selection for random survival forests using maximally selected rank statistics. Statistics in medicine. 2017;36(8):1272–1284.

30. Ishwaran H, Kogalur UB. randomForestSRC: Random Forests for Survival, Regression and Classification (RF-SRC). R package version. 2014;.

31. Wright MN, Ziegler A. ranger: A fast implementation of random forests for high dimensional data in C++ and R. Journal of Statistical Software. 2017;77(i01).

32. Cox DR. Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological). 1972;34(2):187–202.

33. Harrell Jr FE, Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. Statistics in medicine. 1984;3(2):143–152.

34. G¨onen M, Heller G. Concordance probability and discriminatory power in proportional hazards regression. Biometrika. 2005;92(4):965–970.

35. Pencina MJ, D'Agostino RB. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. Statistics in medicine. 2004;23(13):2109–2123.

36. Santos MY, e Sa´ JO, Andrade C, Lima FV, Costa E, Costa C, et al. A big data system supporting bosch braga industry 4.0 strategy. International Journal of Information Management. 2017;37(6):750–760.

37. Schwarz DF, K¨onig IR, Ziegler A. On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data. Bioinformatics. 2010;26(14):1752–1758.

38. Jones Z, Linder F. Exploratory data analysis using random forests. In: Prepared for the 73rd annual MPSA conference; 2015.

39. Ishwaran H, et al. Variable importance in binary regression trees and forests. Electronic Journal of Statistics. 2007;1:519–537.

40. Ishwaran H, Kogalur UB, Gorodeski EZ, Minn AJ, Lauer MS. High-dimensional variable selection for survival data. Journal of the American Statistical Association. 2010;105(489):205–217.

41. Strobl C, Boulesteix A, Zeileis A, Hothorn T. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC bioinformatics. 2007;8(1):25.

42. Wright MN, Ziegler A, K¨onig IR. Do little interactions get lost in dark random forests? BMC bioinformatics. 2016;17(1):145.

43. Breiman L. Random forests. Machine learning. 2001;.

44. Strobl C, Boulesteix A, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. BMC Bioinformatics. 2008;9(1):307.

45. Rutstein S, Winter R, Staveteig S, Yourkavitch J. Urban Child Poverty, Health, and Survival in Low-and Middle-income Countries. In: PAA 2017 Annual Meeting; 2017.

46. Kimani-Murage EW, Fotso JC, Egondi T, Abuya B, Elungata P, Ziraba A, et al. Trends in childhood mortality in Kenya: the urban advantage has seemingly been wiped out. Health & place. 2014;29:95–103.

47. Sousa A, Hill K, Dal Poz MR. Sub-national assessment of inequality trends in neonatal and child mortality in Brazil. International journal for equity in health. 2010;9(1):21.

48. Kunst AE, Mackenbach JP. The size of mortality differences associated with educational level in nine industrialized countries. American journal of public health. 1994;84(6):932–937.

49. Costa JC, da Silva ICM, Victora CG. Gender bias in under-five mortality in low/middle-income countries. BMJ global health. 2017;2(2):e000350.

50. Krishnan A. Gender inequity in child survival: travails of the girl child in rural north India. Ume˚a universitet; 2013.

51. Morgan JN, Sonquist JA. Problems in the analysis of survey data, and a proposal. Journal of the American statistical association. 1963;58(302):415–434.

52. Breiman L, Friedman J, Olshen R, Stone C. Classification and regression trees–crc press. Boca Raton, Florida. 1984;.

53. Gordon L, Olshen R. Tree-structured survival analysis. Cancer treatment reports. 1985;69(10):1065–1069.

54. Bou-Hamad I, Larocque D, Ben-Ameur H, et al. A review of survival trees. Statistics Surveys. 2011;5:44–71.

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

56. Adegbosin, Adeyinka Emmanuel and Stantic, Bela and Sun, Jing. Efficacy of deep learning methods for predicting under-five mortality in 34 low-income and middle-income countries. British Medical Journal Publishing Group. 2020;10(8):e034524,

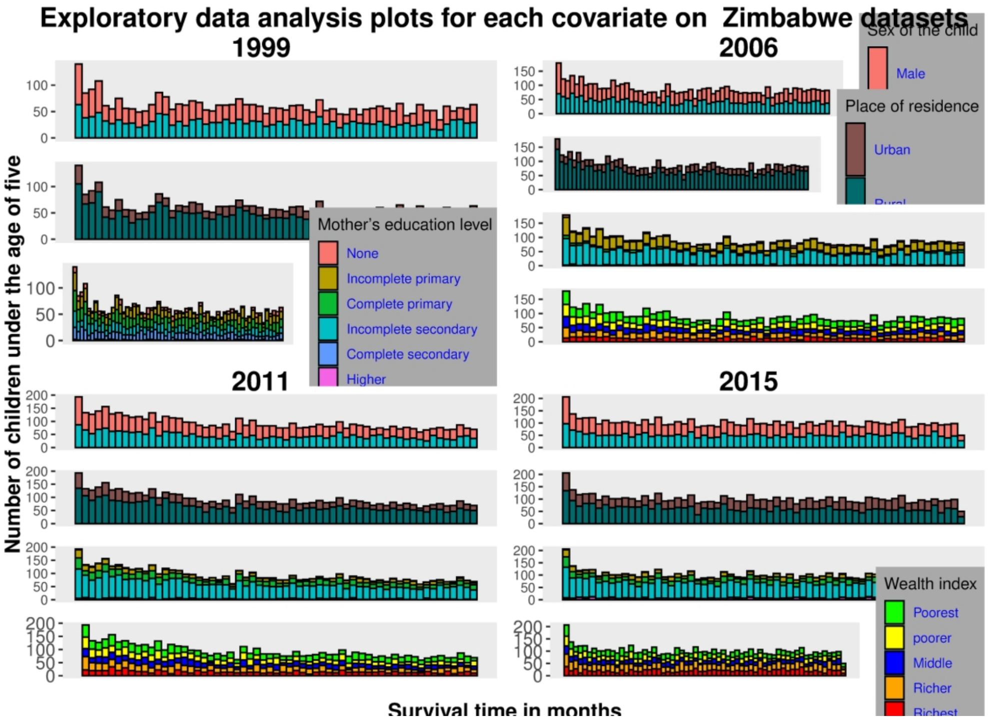
57. Dietterichl TG. Ensemble learning. The handbook of brain theory and neural networks. 2002;.

58. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. The Annals of Applied Statistics. 2008; p. 841–860.

59. Strasser H, Weber C. On the asymptotic theory of permutation statistics. 1999;8:220–250.

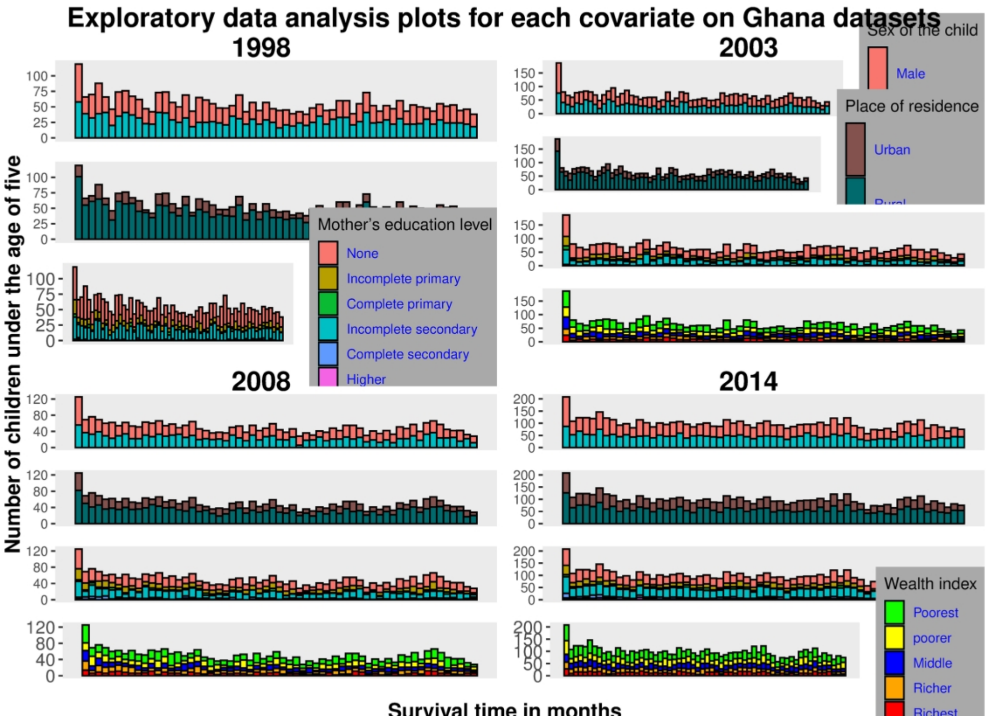
60. R Core Team. R: A Language and Environment for Statistical Computing. https://www.R-project.org/. R Foundation for Statistical Computing.

61. Ishwaran, H and Kogalur, U. B. Package 'randomSurvivalForest'. 2013.



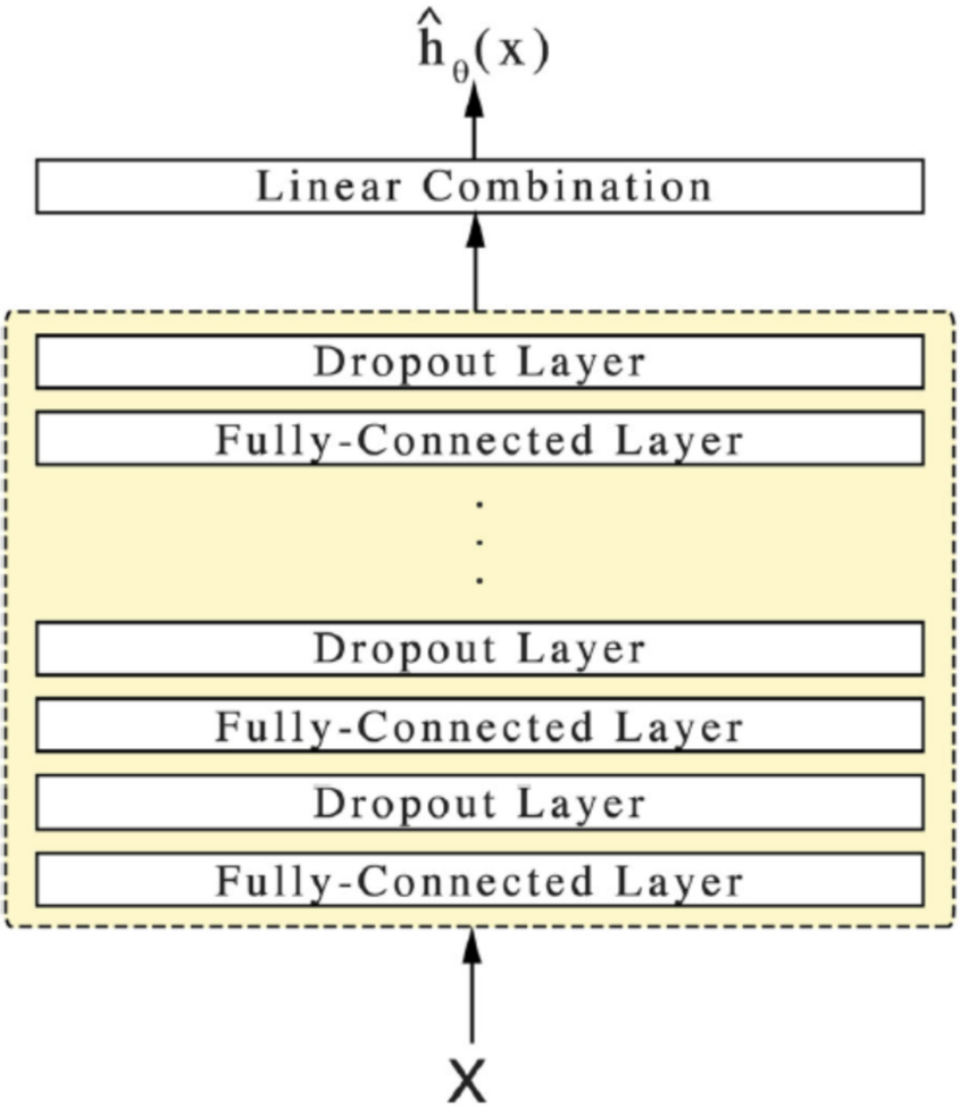
EDA plots for the covariates. Bars indicate number of children under five years for each covariate. Colors correspond to class membership within each covariate

190x137mm (300 x 300 DPI)



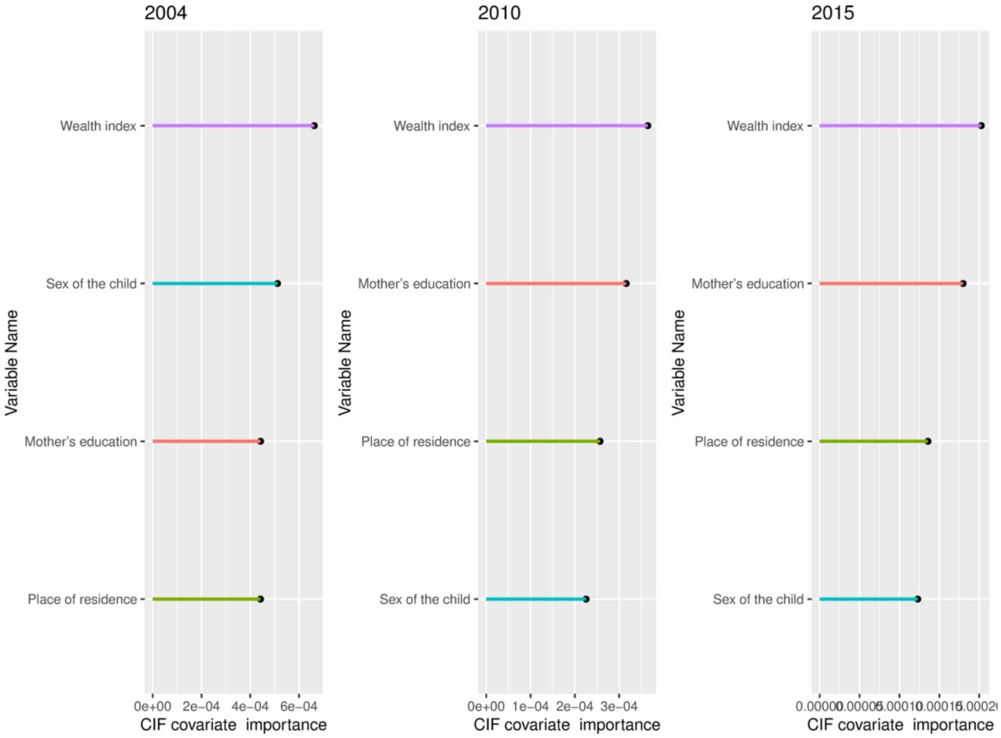
EDA plots for the covariates. Bars indicate number of children under five years for each covariate. Colors correspond to class membership within each covariate

190x137mm (300 x 300 DPI)



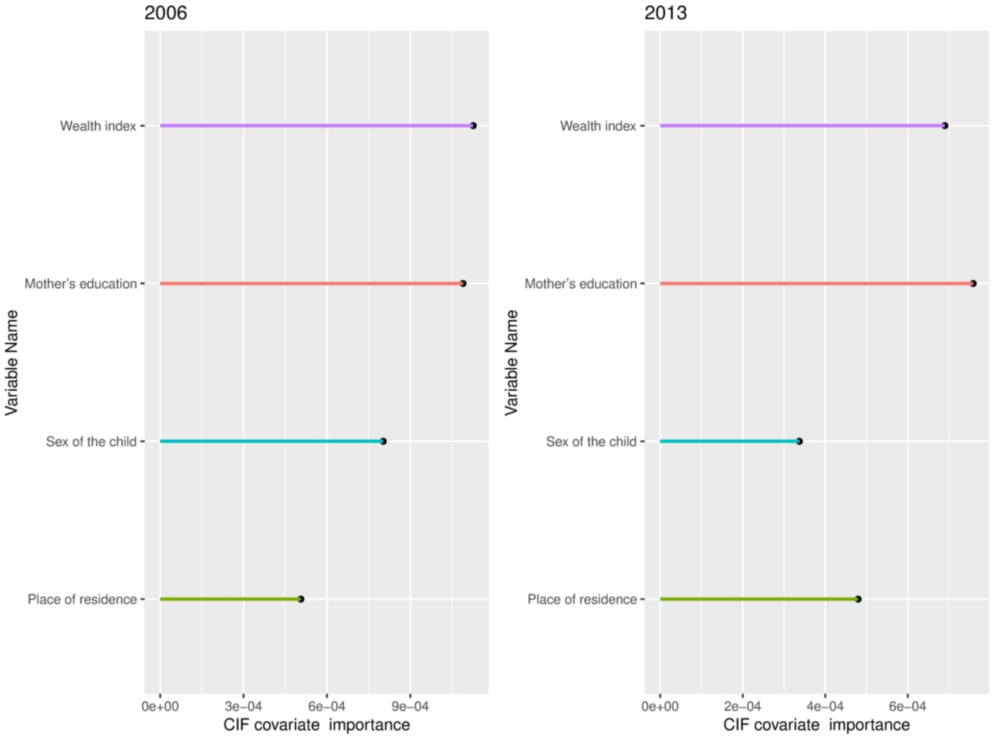
DeepSurv architecture

75x84mm (300 x 300 DPI)



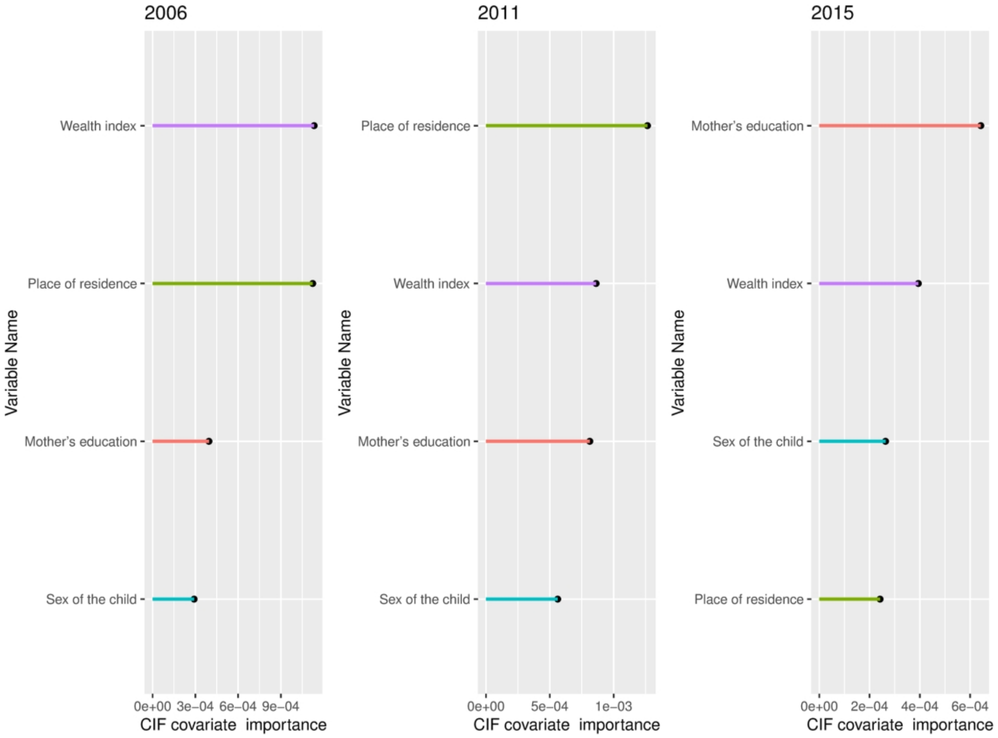
Ranks of importance for the four social economic factors in predicting U5MR in Malawi over the period of 11 years

190x141mm (300 x 300 DPI)



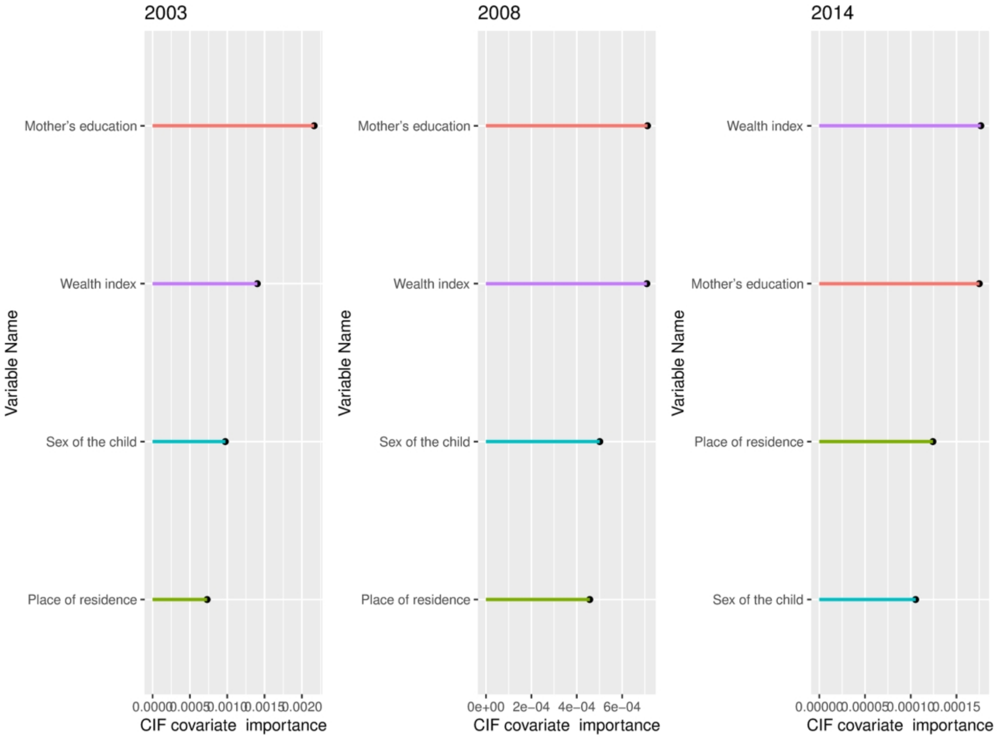
Ranks of importance for the four social economic factors in predicting U5MR in Namibia over the period of 7 years

190x141mm (300 x 300 DPI)



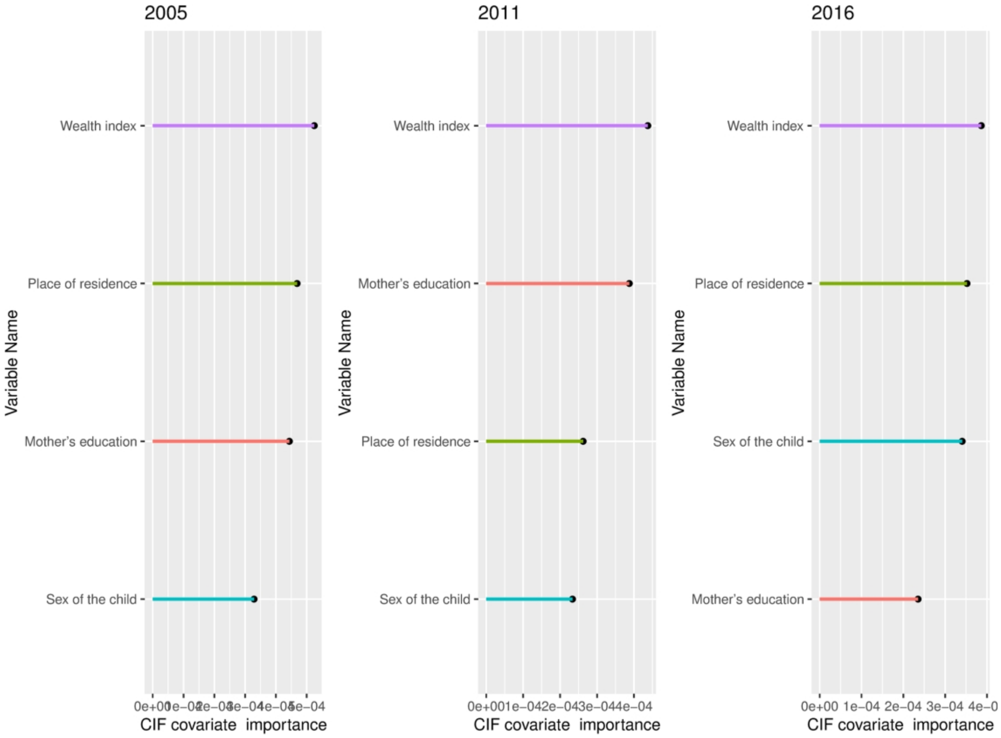
Ranks of importance for the four social economic factors in predicting U5MR in Zimbabwe over the period of 9 years

190x141mm (300 x 300 DPI)



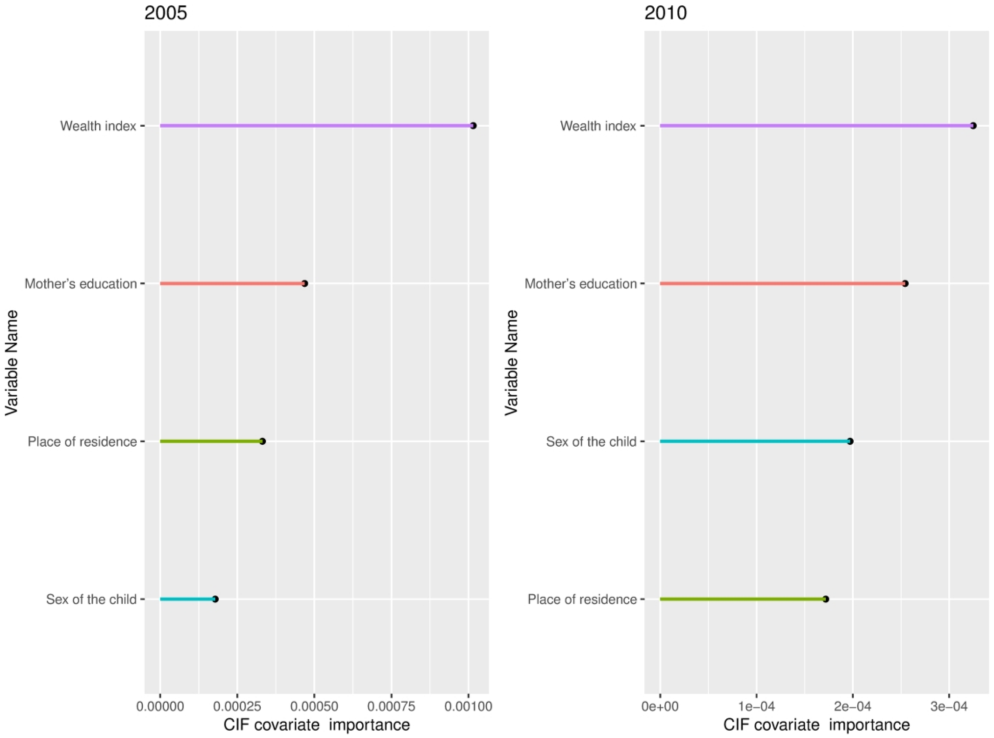
Ranks of importance for the four social economic factors in predicting U5MR in Kenya over the period of 11 years

190x141mm (300 x 300 DPI)



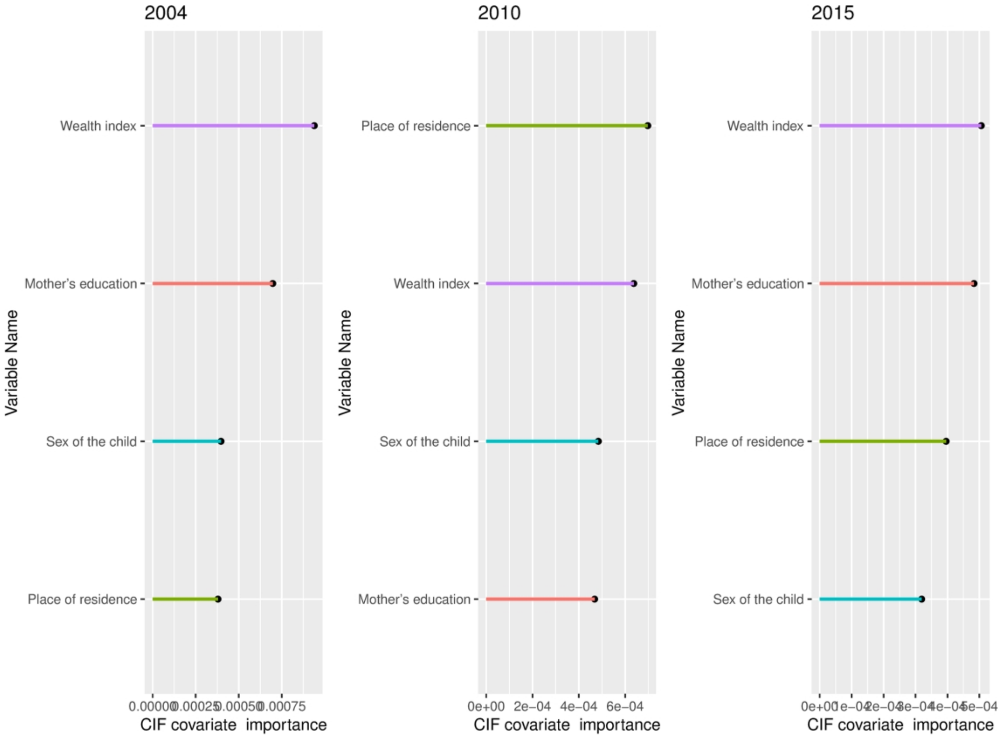
Ranks of importance for the four social economic factors in predicting U5MR in Ethiopia over the period of 11 years

190x141mm (300 x 300 DPI)



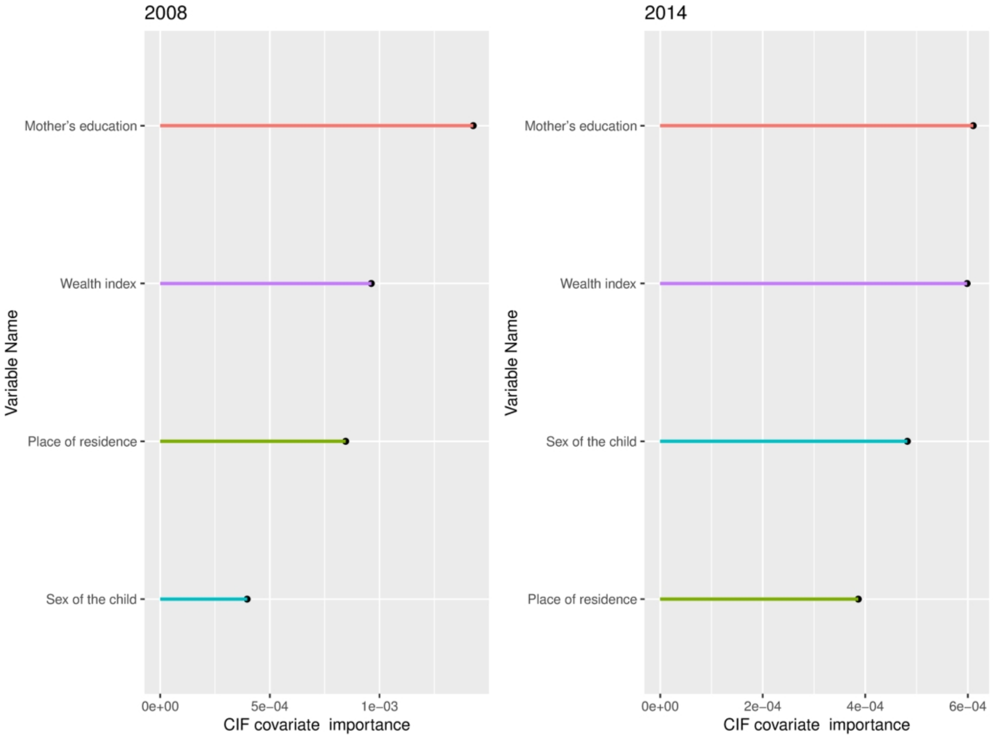
Ranks of importance for the four social economic factors in predicting U5MR in Senegal over the period of 5 years

190x141mm (300 x 300 DPI)



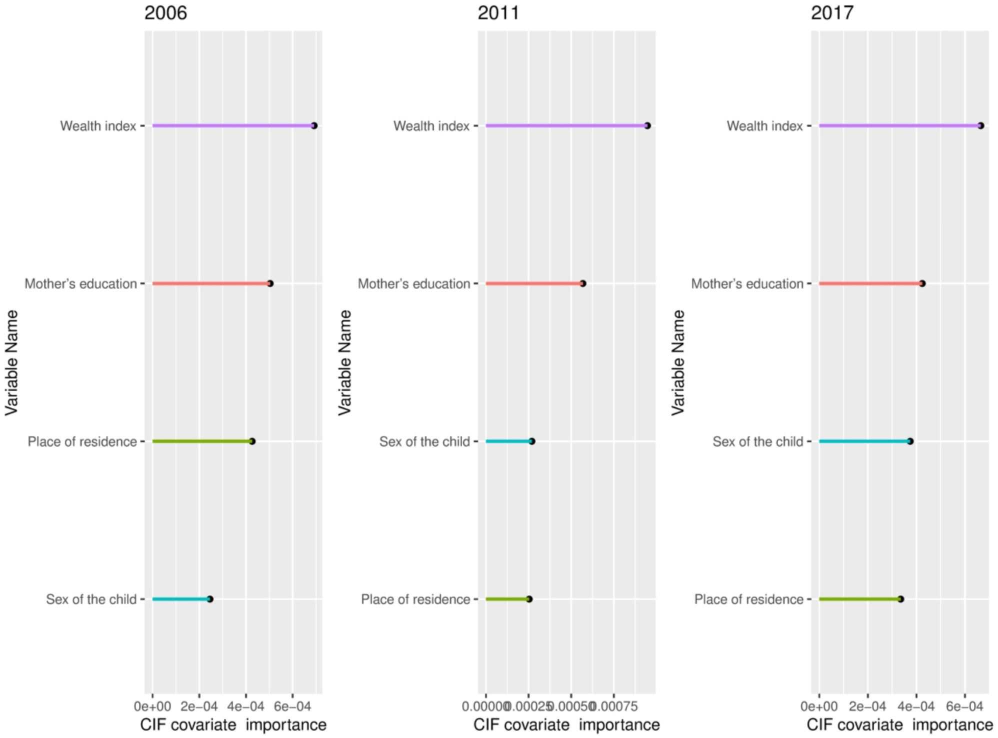
Ranks of importance for the four social economic factors in predicting U5MR in Tanzania over the period of 11 years

190x141mm (300 x 300 DPI)



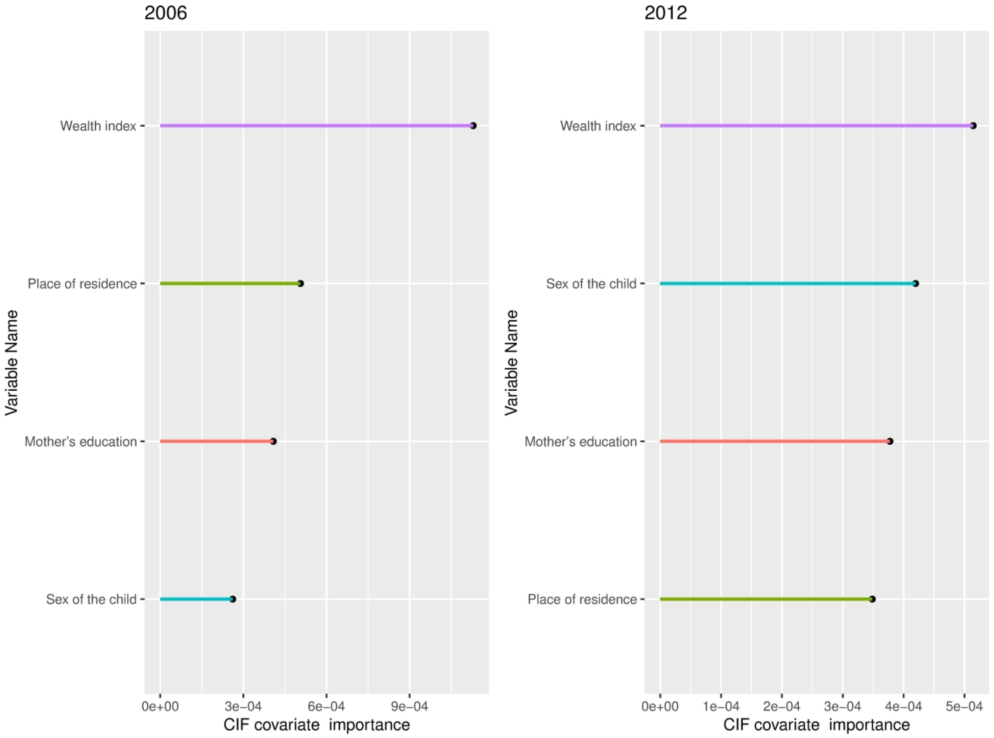
Ranks of importance for the four social economic factors in predicting U5MR in Ghana over the period of 6 years

190x141mm (300 x 300 DPI)



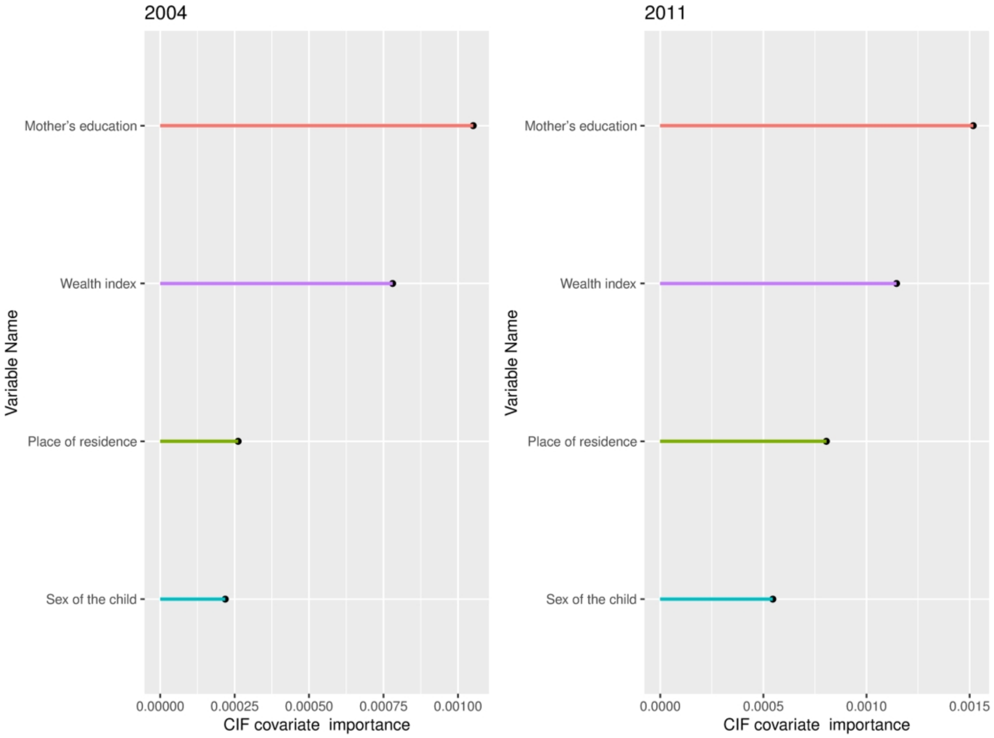
Ranks of importance for the four social economic factors in predicting U5MR in Benin over the period of 11 years

190x141mm (300 x 300 DPI)



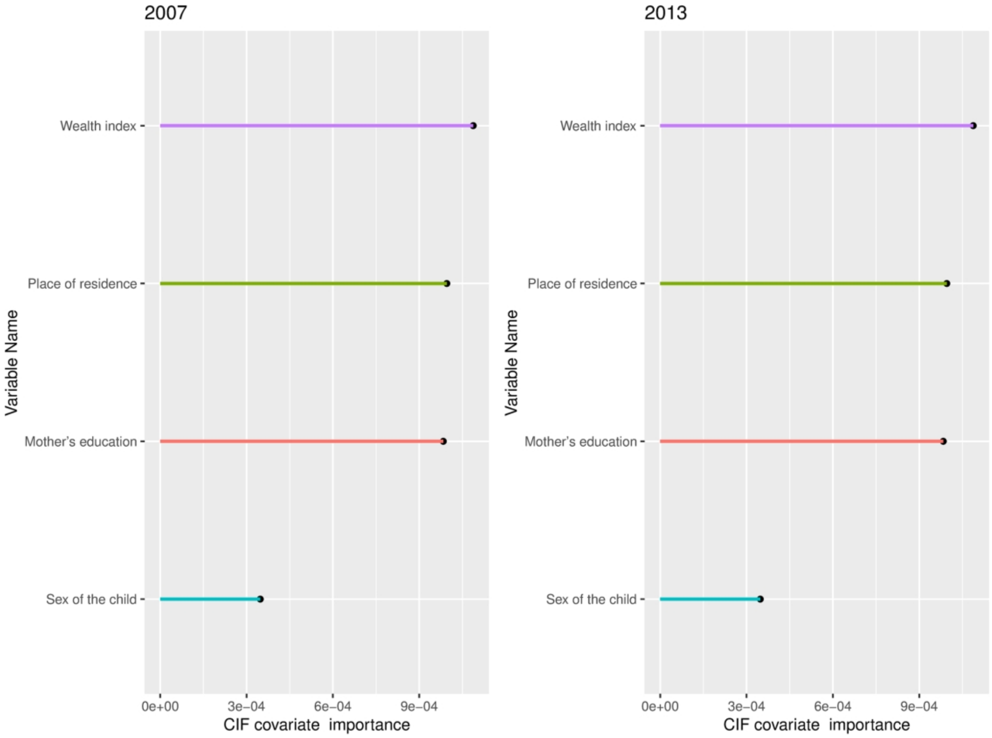
Ranks of importance for the four social economic factors in predicting U5MR in Mali over the period of 6 years

190x141mm (300 x 300 DPI)



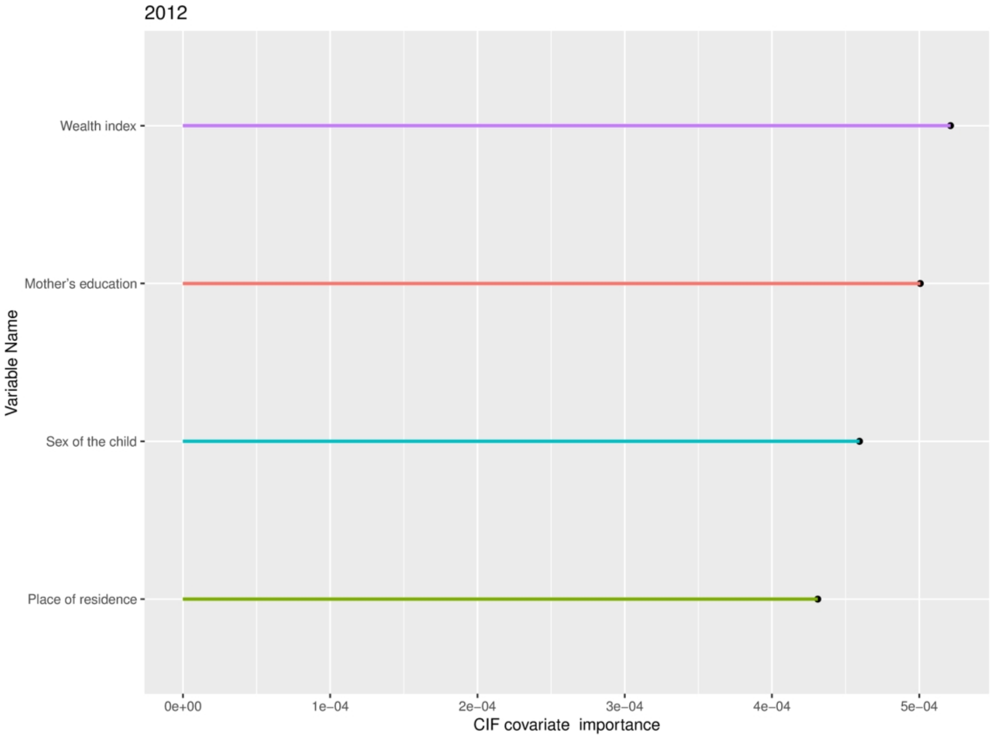
Ranks of importance for the four social economic factors in predicting U5MR in Cameroon over the period of 7 years

190x141mm (300 x 300 DPI)



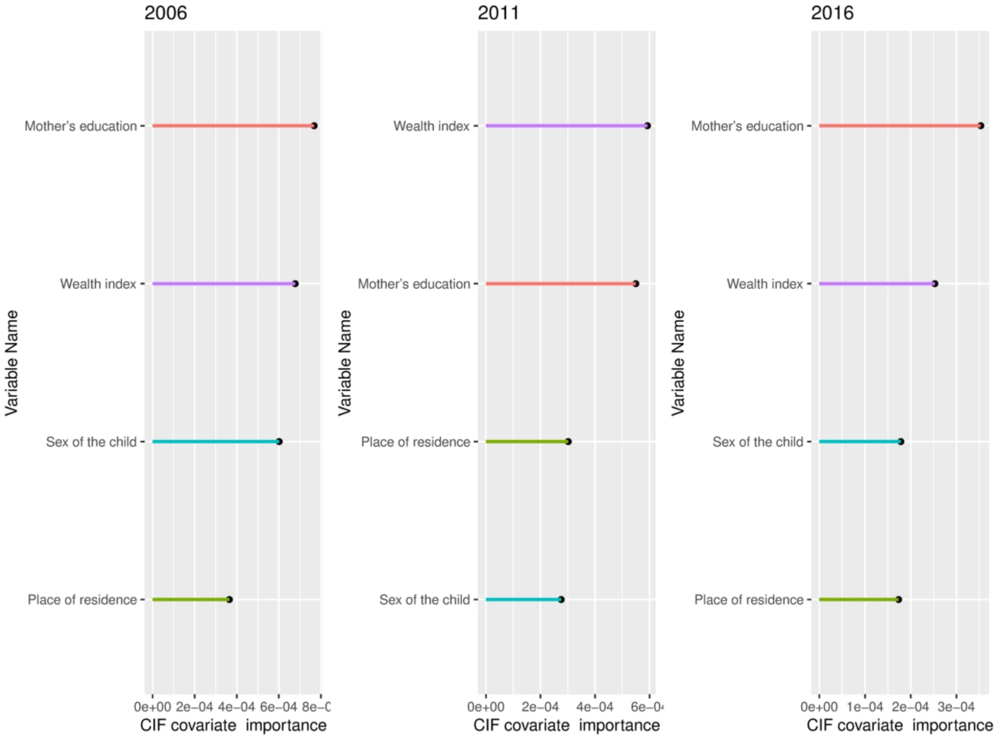
Ranks of importance for the four social economic factors in predicting U5MR in DRC over the period of 6 years

190x141mm (300 x 300 DPI)



Ranks of importance for the four social economic factors in predicting U5MR in Gabon

190x141mm (300 x 300 DPI)



Ranks of importance for the four social economic factors in predicting U5MR in Uganda over the period of 10 years

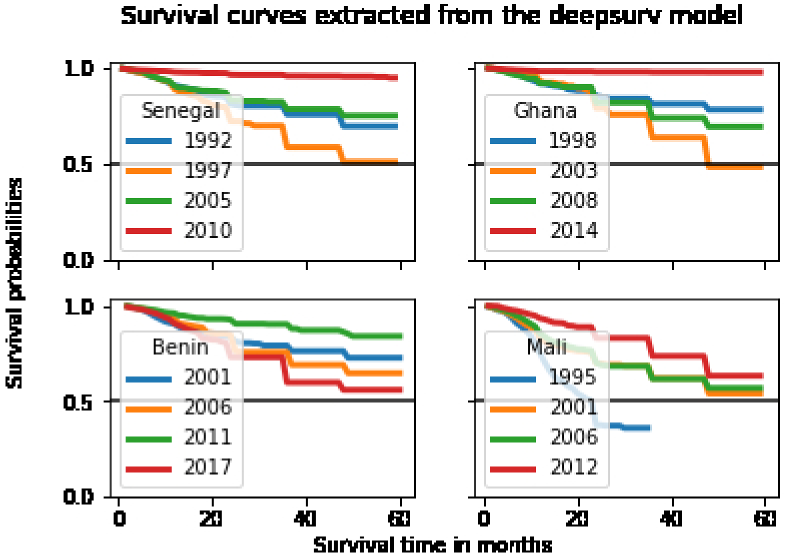
190x141mm (300 x 300 DPI)

Ranks of importance for the four social economic factors in predicting U5MR in Zambia over the period of 6 years

Ranks of importance for the four social economic factors in predicting U5MR in Chad over the period of 10 years

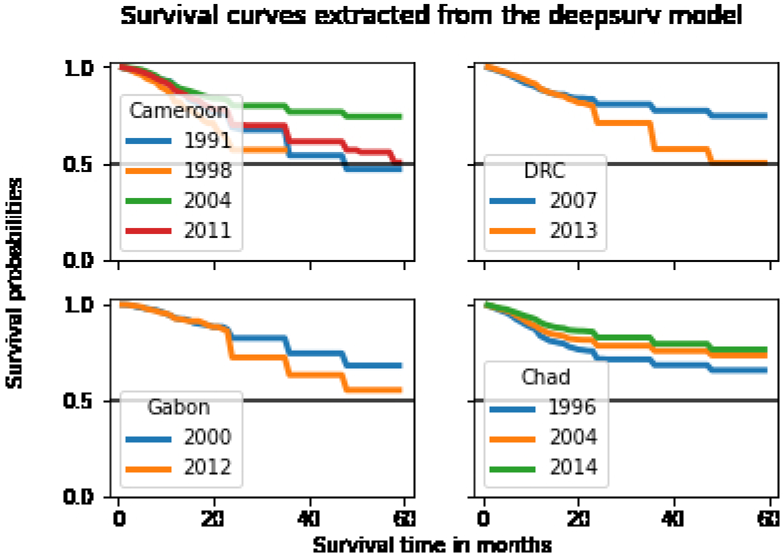
Survival probabilities for the children in the test dataset on the Southern Africa datasets obtained from the deepsurv model

Survival probabilities for some of the children on the Eastern Africa datasets
obtained from the deepsurv model



Survival probabilities for the children in test dataset on the Western Africa datasets obtained from the deepsurv model

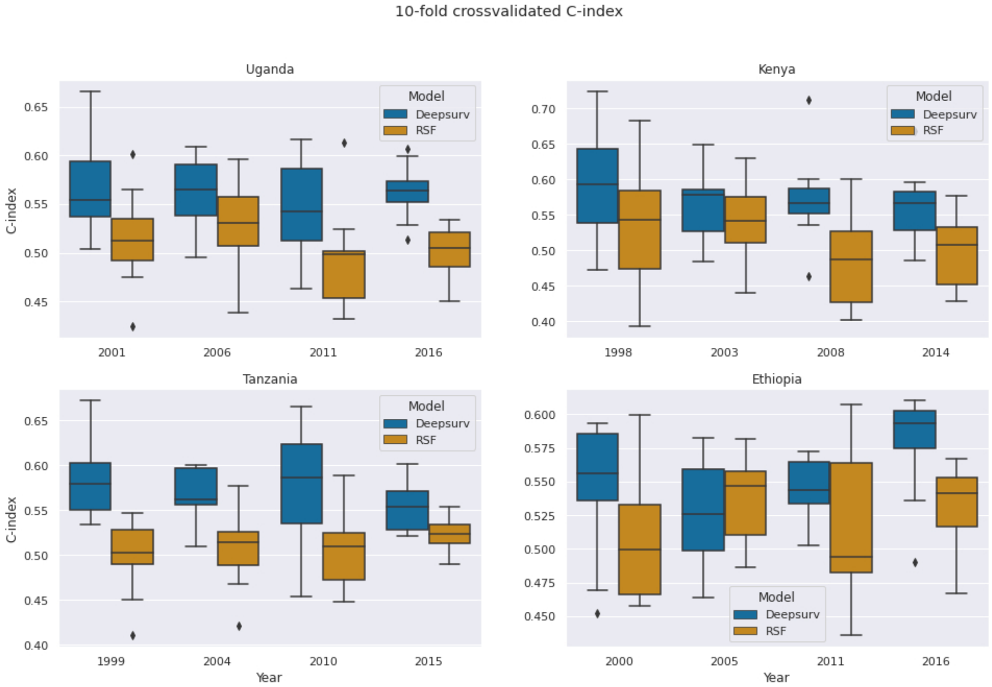
361x263mm (28 x 28 DPI)



Survival probabilities for the children in the test dataset on the Central Africa datasets obtained from the deepsurv model

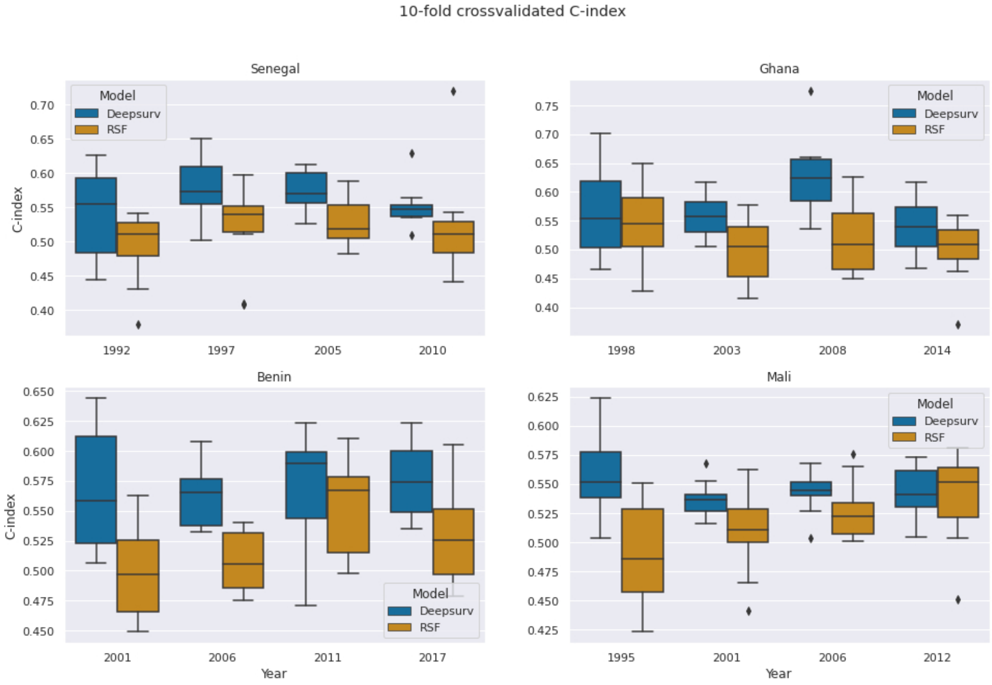
363x263mm (28 x 28 DPI)

Comparison of predictive performance of the deep survival neural network and the random forest models on Southern Africa datasets



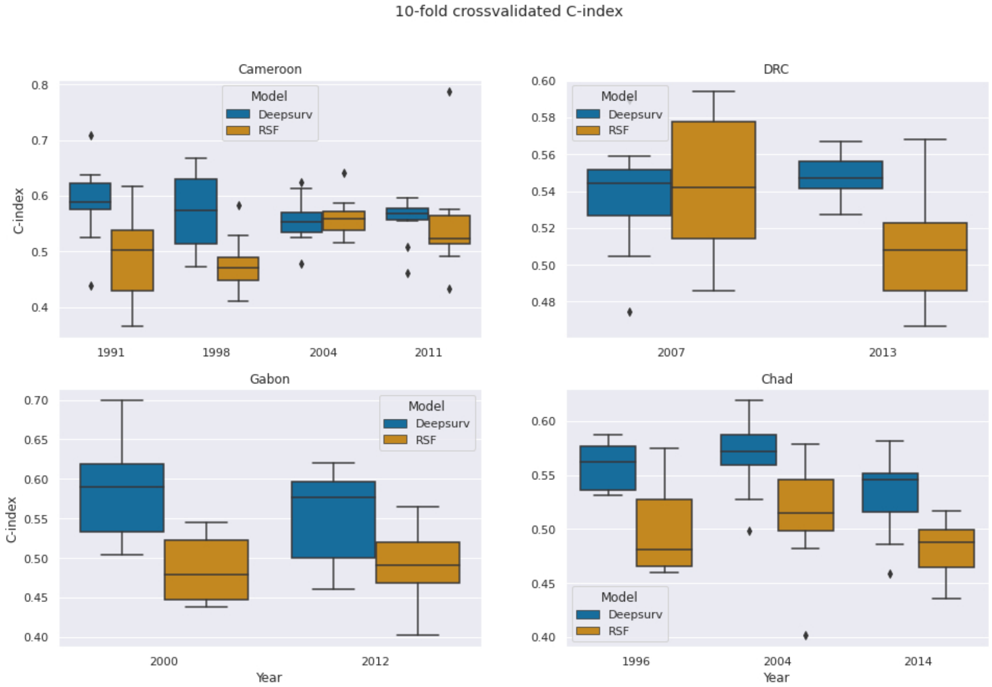
Comparison of predictive performance of the deep survival neural network and the random forest models on Eastern Africa datasets

869x604mm (28 x 28 DPI)



Comparison of predictive performance of the deep survival neural network and the random forest models on Western Africa datasets

875x604mm (28 x 28 DPI)



Comparison of predictive performance of the deep survival neural network and the random forest models on Central Africa datasets

869x604mm (28 x 28 DPI)

Table 2: The total number of deaths under the age of five in each dataset

Zimbabwe	Sex of the child		place of residence		Mother's education level						Wealth index						
	Male	Female	Urban	Rural	None	Incomplete Primary	Complete Primary	Incomplete-Secondary	Complete-Secondary	Higher	Poorest	Poorer	Middle	Richer	Richest	N	Dead
2006	2636	2610	1340	3906	206	1696	330	2870	22	122	1351	1166	958	1019	752	5246	371
2011	2812	2751	1611	3952	100	710	1131	3417	54	151	1366	1145	1001	1178	873	5563	360
2015	3024	3108	2316	3816	63	736	1070	3823	78	362	1244	1075	958	1603	1252	6132	325
Malawi																	
2004	5523	5391	1137	9777	2870	6058	909	709	338	30	2112	2507	2588	2154	1553	10914	1056
2010	9979	9988	1896	18071	3372	12026	1839	1930	693	107	4534	4471	4510	3785	2667	19967	1607
2015	8687	8599	2766	14520	2161	9832	1624	2480	903	286	3909	3743	3369	3191	3074	17286	824
Zambia																	
2007	3181	3220	2073	4328	844	2685	1312	1200	215	145	1385	1390	1467	1355	804	6401	557
2013	6828	6629	4998	8459	1509	5361	2120	3231	750	475	3199	3215	3064	2282	1697	13457	743
Namibia																	
2006	2658	2510	1972	3196	635	1174	410	2228	506	215	1076	1009	1329	1125	629	5168	310
2013	2498	2548	2290	2756	424	864	334	2469	715	240	1089	1113	1121	1058	665	5046	228
Uganda																	
2006	4145	4224	917	7452	2034	4346	835	932	27	195	2139	1820	1555	1491	1364	8369	776
2011	3944	3934	1682	6196	1427	3789	898	1361	84	319	2030	1550	1405	1230	1663	7878	523
2016	7844	7678	2811	12711	2080	7568	2137	2767	162	808	4152	3382	2971	2607	2410	15522	812
Kenya																	
2003	3015	2934	1534	4415	1210	1949	1507	494	538	251	1499	1117	1077	937	1319	5949	502
2008	3134	2945	1467	4612	1300	1915	1515	474	550	325	1777	1079	985	985	1253	6079	373
2014	10633	10331	6828	14136	4585	5905	5150	1861	2142	1321	7178	4348	3497	3131	2810	20964	871
Ethiopia																	
2005	5027	4834	1358	8503	7609	1396	152	466	167	71	2529	1846	1837	1672	1977	9861	859
2011	5987	5667	1986	9668	8142	2691	239	292	94	196	3625	2114	1872	1870	2173	11654	846
2016	5483	5158	1974	8667	6838	2444	234	633	101	391	3993	1782	1466	1308	2092	10641	635
Tanzania																	
2004	4290	4274	1472	7092	2404	1457	3983	608	7	105	1876	1758	1717	1871	1342	8564	712
2010	4009	4014	1511	6512	2043	1274	3780	821	82	23	1610	1815	1715	1656	1227	8023	497
2015	5153	5080	2392	7841	2199	1398	4772	849	926	89	2334	2093	1990	2129	1687	10233	520
Cameroon																	
2004	5814	5918	4691	7041	2917	2838	2068	3414	166	329	2506	2752	2531	2199	1744	11732	998
2011	4060	4065	3160	4965	2109	2117	1531	2216	75	77	1925	1687	1896	1472	1145	8125	844
Chad																	
2004	2839	2796	2504	3131	4174	943	119	341	29	29	916	867	762	1011	2079	5635	709
2014	9472	9151	3973	14650	13424	2898	730	1329	165	77	3559	3786	3902	4097	3279	18623	1722
Democratic republic of Congo (DRC)																	
2007	4476	4516	3575	5417	2214	3086	745	2429	428	90	2038	1855	1745	1871	1483	8992	1005
2013	9301	9415	5504	13212	3933	6521	1925	5020	1086	231	4987	4189	3923	3229	2388	18716	1488
Gabon																	
2012	3030	3037	3713	2354	357	1888	557	2991	102	172	2837	1333	820	608	469	6067	320
Senegal																	
2005	5628	5316	3583	7361	8195	1886	249	548	39	27	2617	2767	2711	1664	1185	10944	838
2010	6342	5984	3645	8681	9225	1904	360	748	47	42	3787	3231	2554	1687	1067	12326	693
Ghana																	
2003	1950	1894	1043	2801	1824	595	228	1069	88	40	1285	859	682	539	479	3844	314
2008	1526	1466	1000	1992	1132	561	161	924	149	65	973	656	504	502	357	2992	198
2014	3066	2818	2344	3540	2042	884	325	2055	354	224	1886	1304	1083	883	728	5884	289
Benin																	
2006	8105	7970	5713	10362	12226	2521	236	980	41	71	3804	3370	3368	3143	2390	16075	1393
2011	6902	6505	4937	8470	9950	1661	420	1180	98	98	3146	2909	2831	2536	1985	13407	728
2017	6910	6679	5401	8188	8936	2252	195	1948	62	196	3020	2776	2670	2639	2484	13589	938
Mali																	
2006	7192	7046	4194	10044	12075	1224	206	651	44	38	2708	2979	3096	3026	2429	14238	1801
2012	5324	5002	2525	7801	8484	733	211	750	72	76	2032	2088	2075	1990	2141	10326	744

COREQ (COnsolidated criteria for REporting Qualitative research) Checklist

A checklist of items that should be included in reports of qualitative research. You must report the page number in your manuscript where you consider each of the items listed in this checklist. If you have not included this information, either revise your manuscript accordingly before submitting or note N/A.

Topic	Item No.	Guide Questions/Description	Reported on Page No.
Domain 1: Research team and reflexivity			
Personal characteristics			
Interviewer/facilitator	1	Which author/s conducted the interview or focus group?	DHS-PAGE3
Credentials	2	What were the researcher's credentials? E.g. PhD, MD	DHS-PAGE16
Occupation	3	What was their occupation at the time of the study?	N/A
Gender	4	Was the researcher male or female?	BOTH MALE AN
Experience and training	5	What experience or training did the researcher have?	DHS-PAGE 16
Relationship with participants			
Relationship established	6	Was a relationship established prior to study commencement?	N/A
Participant knowledge of the interviewer	7	What did the participants know about the researcher? e.g. personal goals, reasons for doing the research	DHS-PAGE 16
Interviewer characteristics	8	What characteristics were reported about the inter viewer/facilitator? e.g. Bias, assumptions, reasons and interests in the research topic	DHS-PAGE16
Domain 2: Study design			
Theoretical framework			
Methodological orientation and Theory	9	What methodological orientation was stated to underpin the study? e.g. grounded theory, discourse analysis, ethnography, phenomenology, content analysis	PAGE 2
Participant selection			
Sampling	10	How were participants selected? e.g. purposive, convenience, consecutive, snowball	DHS-PAGE16
Method of approach	11	How were participants approached? e.g. face-to-face, telephone, mail, email	DHS-PAGE 16
Sample size	12	How many participants were in the study?	TABLE 2-PAGE
Non-participation	13	How many people refused to participate or dropped out? Reasons?	DHS-PAGE 16
Setting			
Setting of data collection	14	Where was the data collected? e.g. home, clinic, workplace	DHS-PAGE 16
Presence of non-participants	15	Was anyone else present besides the participants and researchers?	DHS-PAGE16
Description of sample	16	What are the important characteristics of the sample? e.g. demographic data, date	PAGE 3
Data collection			
Interview guide	17	Were questions, prompts, guides provided by the authors? Was it pilot tested?	DHS-PAGE 16
Repeat interviews	18	Were repeat inter views carried out? If yes, how many?	N/A
Audio/visual recording	19	Did the research use audio or visual recording to collect the data?	NO
Field notes	20	Were field notes made during and/or after the inter view or focus group?	DHS-PAGE 16
Duration	21	What was the duration of the inter views or focus group?	N/A
Data saturation	22	Was data saturation discussed?	N/A
Transcripts returned	23	Were transcripts returned to participants for comment and/or	N/A

Topic	Item No.	Guide Questions/Description	Reported on Page No.
		correction?	
Domain 3: analysis and findings			
Data analysis			
Number of data coders	24	How many data coders coded the data?	ONE
Description of the coding tree	25	Did authors provide a description of the coding tree?	NO
Derivation of themes	26	Were themes identified in advance or derived from the data?	DERIVED FROM
Software	27	What software, if applicable, was used to manage the data?	R AND PYTHON
Participant checking	28	Did participants provide feedback on the findings?	NO
Reporting			
Quotations presented	29	Were participant quotations presented to illustrate the themes/findings? Was each quotation identified? e.g. participant number	NO
Data and findings consistent	30	Was there consistency between the data presented and the findings?	YES
Clarity of major themes	31	Were major themes clearly presented in the findings?	YES
Clarity of minor themes	32	Is there a description of diverse cases or discussion of minor themes?	YES-PAGE 15

Developed from: Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. *International Journal for Quality in Health Care*. 2007. Volume 19, Number 6: pp. 349 – 357

Once you have completed this checklist, please save a copy and upload it as part of your submission. DO NOT include this checklist as part of the main manuscript document. It must be uploaded as a separate file.

Standards for Reporting Qualitative Research (SRQR)*

http://www.equator-network.org/reporting-guidelines/srqr/

Title and abstract

	Page/line no(s).
Title - Concise description of the nature and topic of the study Identifying the study as qualitative or indicating the approach (e.g., ethnography, grounded theory) or data collection methods (e.g., interview, focus group) is recommended	PAGE 1
Abstract - Summary of key elements of the study using the abstract format of the intended publication; typically includes background, purpose, methods, results, and conclusions	PAGE 1

Introduction

Problem formulation - Description and significance of the problem/phenomenon studied; review of relevant theory and empirical work; problem statement	PAGE 2
Purpose or research question - Purpose of the study and specific objectives or questions	PAGE 3

Methods

Qualitative approach and research paradigm - Qualitative approach (e.g., ethnography, grounded theory, case study, phenomenology, narrative research) and guiding theory if appropriate; identifying the research paradigm (e.g., postpositivist, constructivist/ interpretivist) is also recommended; rationale**	PAGE 6 TO 10
Researcher characteristics and reflexivity - Researchers' characteristics that may influence the research, including personal attributes, qualifications/experience, relationship with participants, assumptions, and/or presuppositions; potential or actual interaction between researchers' characteristics and the research questions, approach, methods, results, and/or transferability	PAGE 11 TO 14
Context - Setting/site and salient contextual factors; rationale**	
Sampling strategy - How and why research participants, documents, or events were selected; criteria for deciding when no further sampling was necessary (e.g., sampling saturation); rationale**	
Ethical issues pertaining to human subjects - Documentation of approval by an appropriate ethics review board and participant consent, or explanation for lack thereof; other confidentiality and data security issues	PAGE 16
Data collection methods - Types of data collected; details of data collection procedures including (as appropriate) start and stop dates of data collection and analysis, iterative process, triangulation of sources/methods, and modification of procedures in response to evolving study findings; rationale**	PAGE 3 TO 5 AND PAGE 16

Data collection instruments and technologies - Description of instruments (e.g., interview guides, questionnaires) and devices (e.g., audio recorders) used for data collection; if/how the instrument(s) changed over the course of the study	N/A
Units of study - Number and relevant characteristics of participants, documents, or events included in the study; level of participation (could be reported in results)	PAGE 16
Data processing - Methods for processing data prior to and during analysis, including transcription, data entry, data management and security, verification of data integrity, data coding, and anonymization/de-identification of excerpts	N/A
Data analysis - Process by which inferences, themes, etc., were identified and developed, including the researchers involved in data analysis; usually references a specific paradigm or approach; rationale**	PAGE 11 TO 14
Techniques to enhance trustworthiness - Techniques to enhance trustworthiness and credibility of data analysis (e.g., member checking, audit trail, triangulation); rationale**	PAGE 17

Results/findings

Synthesis and interpretation - Main findings (e.g., interpretations, inferences, and themes); might include development of a theory or model, or integration with prior research or theory	PAGE 15 TO 16
Links to empirical data - Evidence (e.g., quotes, field notes, text excerpts, photographs) to substantiate analytic findings	PAGE 15 TO 16

Discussion

Integration with prior work, implications, transferability, and contribution(s) to the field - Short summary of main findings; explanation of how findings and conclusions connect to, support, elaborate on, or challenge conclusions of earlier scholarship; discussion of scope of application/generalizability; identification of unique contribution(s) to scholarship in a discipline or field	PAGE 15
Limitations - Trustworthiness and limitations of findings	PAGE 1

Other

Conflicts of interest - Potential sources of influence or perceived influence on study conduct and conclusions; how these were managed	PAGE 17
Funding - Sources of funding and other support; role of funders in data collection, interpretation, and reporting	PAGE 17

*The authors created the SRQR by searching the literature to identify guidelines, reporting standards, and critical appraisal criteria for qualitative research; reviewing the reference lists of retrieved sources; and contacting experts to gain feedback. The SRQR aims to improve the transparency of all aspects of qualitative research by providing clear standards for reporting qualitative research.

**The rationale should briefly discuss the justification for choosing that theory, approach, method, or technique rather than other options available, the assumptions and limitations implicit in those choices, and how those choices influence study conclusions and transferability. As appropriate, the rationale for several items might be discussed together.

Reference:

O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. **Standards for reporting qualitative research: a synthesis of recommendations.** *Academic Medicine*, Vol. 89, No. 9 / Sept 2014 DOI: 10.1097/ACM.0000000000000388

BMJ Open

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

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2021-049786.R1
Article Type:	Original research
Date Submitted by the Author:	04-Sep-2021
Complete List of Authors:	Nasejje, Justine B; University of the Witwatersrand, Statistics and Actuarial Science Mbuvha, Rendani; University of the Witwatersrand, Statistics and Actuarial Science Mwambi, Henry; University of Kwazulu-Natal, School of Mathematics, Statistics and Computer Science
<b>Primary Subject Heading</b>:	Public health
Secondary Subject Heading:	Health policy, Global health, Public health
Keywords:	Health policy < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, PUBLIC HEALTH, Community child health < PAEDIATRICS

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

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

1 School of Statistics and Actuarial science, University of Witwatersrand, Jan Smuts Avenue, Johannesburg, Gauteng, South Africa

2 School of Statistics, Mathematics and Computer Science, University of KwaZulu-Natal, King Edward Avenue, Pietermaritzburg, South Africa

* Corresponding author E-mail: justine.nasejje@wits.ac.za

Abstract

Objectives We use 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-five mortality rate (U5MR) in sub-Saharan Africa have evolved.

Settings This work consist of multiple cross-sectional studies. We analysed data from the Demographic Health Surveys (DHS).

Participants A total of n= 85,688 children from eleven datasets drawn from four countries each representing a sub-region in sub-Saharan Africa was analysed.

Primary and secondary outcomes The primary outcome variable is U5MR; the secondary outcomes to obtain the ranks of importance of the four socioeconomic factors over-time and comparing 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 out of the eleven datasets. Wealth index ranked first in three and second in eight out of the eleven datasets. Place of residence ranked first in two out of the eleven datasets. Based on these rankings, the mother's education and wealth index are the most dominant factors. The four factors showed a favourable survival outcome over-time confirming that the past interventions aimed at targeting these factors are yielding positive results. The DeepSurv model has a higher predictive performance with mean concordance indexes (between 67% to 80%), above 50% compared to the RSF model.

Conclusions The study reveals that children under the age of five in sub-Saharan Africa have favourable survival outcomes associated with the four socioeconomic factors over-time. It also shows that deep learning 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 (SDG3).

Strengths and limitations of the study

- The study used machine learning methods which when compared to classical statistical models are very flexible.

- Machine learning methods have fewer assumptions and are adapted to fitting 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 U5MR was the fourth Millennium Development Goals (MDGs) drafted in the year 2000, and the world sprang into action to achieve it and now it appears within the third Sustainable Development Goal (SDG3).

The probability of a child dying before the age of 5 years (U5MR) is a global indicator of societal and national development as it serves as a key marker of health equity and access.[1] The fourth Millennium Development Goal (MDG4) which previously stated that, reducing under-five mortality by two-thirds in the period between 1990 and 2015 now appears in the third Sustainable Development Goal (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 still exists substantial inequalities between subgroups of the population within countries.[2-3] These sub-groups are based on factors such as, wealth index, maternal factors such as education level, place of residence, sex of the child, among others. The Mosley and Chen framework,[4] categorizes these socio-economic factors as the distal determinants of child mortality.[4]

Classical statistical parametric regression models such as the logistic regression model, semi-parametric models like the Cox proportional hazard models (CPH) and generalized additive models have been widely used to study determinants of U5MR .[1, 5-11] A study by Sahu et al.,[7] 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 socio-demographic and economic factors were observed over the period. In a study by Sahu et al.,[7] it was concluded that mother’s education level and sex of the child were among the factors responsible for 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.[12] Mothers’ education, place of residence and sex of the child were also found significant in influencing U5MR trends in Nigeria.[13] Although the CPH and the logistic regression models are very robust, they are often criticised for their restrictive assumptions and hence may lead to bias if care is not taken when preparing the data for analysis.[14] Classical machine learning approaches which include nearest neighbours, neural networks, kernel methods, penalized least squares and data partitioning methods such as decision trees (CART) and random forests are among the alternative approaches to parametric and semi-parametric 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 to the statistical models like the CPH model.[21-25] A recent study by Adegbosin et al.,[25] recommended the use deep learning models in understanding the determinants of U5MR in low- 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 have often been stated among the top predictors of under-five mortality in the Sub-Saharan region. With the launch of the millennium development goals in the year 2000, we saw the convergence of the development agenda of United Nations Development Programme (UNDP); United Nations Environment Programme (UNEP); World health organization (WHO); United Nations Children's Fund (UNICEF); United Nations Educational, Scientific and Cultural Organization (UNESCO); and other development agencies to come up with funding and programmes targeting the inequalities that existed to achieve these goals. [26] Despite the substantial improvement made with the MDG4, inequalities persist till today, and the progress has been uneven. Now that the MDG4 appears in the SDG3 with an even wider age range, we need an evidence-based approach to achieve it by using existing datasets to inform policy.

The study uses two machine learning models; the random survival forest model and the deep survival neural network to answers 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 overtime especially after the expiry of the MDGs? Which of the two machine learning methods, the RSF and the DeepSurv model are effective in predicting U5MR?

Studying how the rank in importance of these factors in determining U5MR has evolved over time can help redirect resources to the right sectors and hence be on-course to achieving SDG3. In this study, therefore we train a random survival forest 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 has evolved over time. The random survival forest 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 extract survival curves to assess whether there is a favourable survival outcome for children under the age of five associated with these factors in this region over-time.

The contributions of this work are as follows: 1) to identifying the importance rankings of the four socioeconomic factors in U5MR prediction in Sub-Saharan Africa 2) to present how the ranking of these factors have changed over time

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 to the four economic factors. These contributions are aimed at assisting policymakers in designing new interventions while also providing evidence of how past interventions have worked through presenting changes in predictive importance rankings of the four socioeconomic factors over-time.

Methods

Data

Datasets of completed Standard Demographic and Health Surveys (DHS) from four countries each selected to represent the four sub-regions (Southern, Central, Eastern and Western Africa) in sub-Saharan Africa are used. DHS funded by USAID, UNFPA, UNICEF, Irish Aid and the United Kingdom government have over the years (since 1988), provided datasets which are 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 program website. The datasets from the DHS program are available on request by a researcher based anywhere in the world. The outcome variable is under-five survival time, and this information was obtained from the birth history of interviewed women aged between 15 to 49 years of age. All the datasets used in this analysis comprised of children dead or alive, born in the period of five years preceding the date of the survey. This is done 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 education, wealth index of the household, and sex of the child. The study randomly selected four countries from each sub-region as shown in Table 1 below. From the four countries, the study considered analysis on only one country per sub-region.

Table 1. The standard Demographic and Health Survey datasets used for this study by region

Southern region				**Eastern Region**			
Zimbabwe	**Malawi**	**Namibia**	**Zambia**	**Uganda**	**Kenya**	**Tanzania**	**Ethiopia**
1999	2000	1992	1996	2001	1998	1999	2000
2006	2004	2000	2001	2006	2003	2004	2005
2011	2010	2006	2007	2011	2008	2010	2011
2015	2015	2013	2013	2016	2014	2015	2016
Western region				**Central region**			
Senegal	**Ghana**	**Benin**	**Mali**	**Cameroon**	**DRC**	**Gabon**	**Chad**
1992	1998	2001	1995	1991	2007	2000	1996
1997	2003	2006	2001	1998	2013	2012	2004
2005	2008	2011	2006	2004			2014
2010	2014	2017	2012	2011			

Table 1. Datasets available for each of the four selected countries in the four regions of sub-Saharan Africa and the year the survey was conducted.

Data pre-processing

Like all survey data, the DHS datasets contain many features or variables. In this study we considered only four features for our analysis, that is place of residence, mother's level of education, wealth index, and sex of the child. All the other features in the datasets were excluded from this analysis. The response variable was calculated differently depending on the survival status of the child. The children under the age of five that were still alive at the time of the survey had their survival time calculated as the difference between the year of the interview and their year of birth. For the children that were dead at the time of the survey, their survival time was calculated as the difference between the year of the interview and the year of death. The response variable was later transformed into months for this analysis. For each of the dataset, a data frame containing the four features, the response variable (survival time in months) and the status indicator (child is dead or alive) was created. We had complete information across all the datasets for the features considered in this analysis. It is also important to note that for some of the datasets that were collected in the 90's and the early 2000's, wealth index was not available as a feature. These datasets were therefore excluded in our final analysis to allow meaningful comparisons. In total we analysed eleven datasets in this study, and these are summarised in the tables below.

Table 2: The total number of children under the age of five per feature category

	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: The total number of children under the age of five per feature category

	Wealth index					Total
	Poorest	**Poorer**	**Middle**	**Richer**	**Richest**	
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**
Total						**85688**

Table 2 and Table 3 give the counts of the number of children under the age of five for each of the feature category in all the datasets considered for analysis. Table 3 shows that the total number of children from the multiple DHS datasets considered for this study is 85,688.

Patient and Public Involvement

There are no patients involved in this study

Models

The CPH model is the most frequently used model to analyse 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 involving 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 random survival forests which can enable someone detect complex relationships in survival datasets 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] Even though they have exhibited a good performance in predicting survival outcomes, there are few studies aimed at understanding factors associated to U5MR that have embraced these methods.[15,27] Recently, with the advancement of the machine learning methods, deep learning methods have also been added to the toolbox of methods to analyse survival data.[21] The fact that most datasets collected have complex structures, using models that have very strict assumptions may lead to bias and hence misleading policy implementations. In this study, we apply two machine learning models on datasets from sub-Saharan Africa. These two models are the random survival forest, and the deep survival neural network model (DeepSurv).[17,21]

Random survival forests

Random survival forests (RSF) 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 methods in addressing the above-mentioned challenges of the CPH model. The algorithm of the random survival forest model by Brieman et al.,[28] is described in detail below but first, we describe the survival tree algorithm an important building block of the forest.

Survival trees

The regression tree algorithm for right censored data is an extension of the CART algorithm by Breiman et al.,[28]. Below is the general algorithm for survival trees.[29-31]

Algorithm1 : 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 ($\alpha$ and $\beta$) 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 i.e., has no less than $d_0 > 0$ unique observed events.
```

An RSF model is a collection of survival trees because a single tree is always not a good probability estimator due to its short comings 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. The algorithm for building an RSF model as presented by Ishwaran et al.,[17] is given below as follows.

Algorithm2 : 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.
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 maximizes 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 deaths.
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.
```

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

In this study we used a special type of survival forest model, known as the conditional inference survival forest model (CIF).[34-35] The CIF has an advantage over the original random survival forest 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 increasingly becoming popular as additional models to the toolbox of models aimed at predicting survival outcomes. They look very promising especially in application to large datasets that could be having many covariates with non-linear effects on the survival outcome. It is important to note that neural networks are only very good for predicting outcomes but not able to give explanations or quantify covariate effects on the outcomes. Initially, a single hidden layer feed-forward neural network was trained to survival data and their 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 a 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 gentle introduction to the CPH model.

Cox proportional hazards model

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

$$\lambda(t, X) = \lambda_0(t)\exp(h(X)), \qquad (1)$$

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 $\hat{h}_{\beta}\,(X) = \beta' X$. The estimates $(\hat{\beta})$ of the parameters (β) are obtained by maximising the partial likelihood. Suppose that there are k distinct event times, and $t_1 < t_2 < .... < t_k$ represent the ordered distinct event times, the partial likelihood is given as

$$L\left(\beta\right) = \prod_{i=1}^{k} \frac{\exp\left(\hat{h}_{\beta}\left(X_i\right)\right)}{\sum_{j\in\Re(t_i)} \exp\left(\hat{h}_{\beta}\left(X_i\right)\right)}. \quad (2)$$

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

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

In this study, however, the Scaled Exponential Linear Unit (SELU) is used as an activation function because of its advantages over the ReLU. ReLUs 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 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 zero. If this happens to many cells in your network, the power of the trained network stays below its theoretical capabilities. It is given as

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

Where $\gamma > 0$ and $\lambda > 0$ 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 $\hat{h}_{\beta}\,(X) = \beta^{0}X$ in equation 2 by the output of $\hat{h}_{\theta}\,(X) = \exp\ (g\ (X, \theta))$ of the neural network. The proportional hazards model becomes

$$h_{\theta}\,(X_i) = \exp\ (g\ (X_i, \theta)). \quad (4)$$

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

$$L\left(\theta\right) = \prod_{i:\delta_i=1} \frac{\exp\left(\hat{h}_{\theta}\left(X_i\right)\right)}{\sum_{j\in\Re(t_i)} \exp\left(\hat{h}_{\theta}\left(X_i\right)\right)}. \quad (5)$$

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

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

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

$$l(\theta) = -\frac{1}{n_{\delta_1}} \sum_{i:\delta_i=1} \left(\hat{h}_\theta (X_i) - log \sum_{j \in \Re(t_i)} \exp\left(\hat{h}_\theta (X_j) \right) \right) + \alpha \|\theta\|_2^2 \quad (7)$$

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 regularization parameter for the L_2 norm. Gradient descent optimization is used to find the weights of the network which minimise the loss function. The DeepSurv neural network architecture adapted for this study is described in detail by Katzman et al.,[21]. The Figure 1 below shows its architecture. It is a deep feed-forward neural network implemented as

Fig 1. DeepSurv architecture Katzman et al.,[21].

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 dropout 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 1**e**-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 $\hat{Y}$, the C-index is given by

$$C = \frac{\sum_{i:\delta_i=1} \sum_{y_i<y_j} I\left(\hat{Y}_i < \hat{Y}_j\right)}{\text{Number of Comparable Pairs}} \quad (8)$$

In survival analysis, shorter survival time means smaller predicted outcome. C-index value of above 0.5 means better agreement among comparable pairs.[41-43]

Over-fitting 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 mostly commonly used methods to evaluate the predictive performance of machine learning models. The test error is known to be very 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 $K-fold$ crossvalidation is recommended. $K-fold$ crossvalidation 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 use 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 combined 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 fold of the 10 folds have been used as the testing set.

Measures of covariate importance

To understand which factors are important in influencing predictions, the random survival forests model has a measure of estimating importance of each covariate. It is generally referred to as the variable importance measure (VIMP).[45-48] Variables are selected because of their importance in predicting the survival outcome. The basic measure of variable importance is by counting 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. For illustration, first consider a tree grown and its prediction accuracy ($\hat{e}$), calculated using the out-of-bag (OOB) observations. Secondly, randomly permute the values of the factor of interest, (X_j) for all individuals. Note that permutation breaks the original relationship of the covariate with the survival outcome. Obtain a new value for prediction accuracy, ($\hat{e}_j$) using OOB observations. Compare $\hat{e}_j$, with $\hat{e}$ of the original classification for covariate, X_j. Calculate, argmax $\{0; \hat{e}_j - \hat{e}\}$. The difference between the accuracy before and after permutation provides the importance of the covariate, X_j 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

We extracted our most important variables in predicting child survival from our datasets using 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 are shown in Figures 2 -5 below. The ranks of feature importance presented here are from one country in each sub-region that was selected to represent it.

Fig 2. Ranks of importance for the four socioeconomic factors in predicting U5MR in Zimbabwe over the period of 9 years

In Figure 2, the two most important predictors of U5MR in Zimbabwe in 20006 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.

Fig 3. Ranks of importance for the four social economic factors in predicting U5MR in Ghana over the period of 10 years

In Figure 3, mother's education is ranked first for the years 2008, 2014 and wealth index second in both datasets. In Figure 4, wealth index and mothers' 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 well as 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.

Fig 4. Ranks of importance for the four social economic factors in predicting U5MR in Uganda over the period of 10 years

Fig 5. Ranks of importance for the four social economic factors in predicting U5MR in Chad over the period of 10 years

Figures 2-5 show that mother's education is ranked first in five out of the eleven dataset and Wealth index ranked first in three out of the eleven datasets but second in eight out of the eleven 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 two out of the eleven datasets and second in one of the eleven datasets making it among the top three predictors of under-five 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. Sex of the child's rank of importance is not anywhere near the top two in all the datasets considered for analysis. In-fact it was ranked fourth 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- and middle-income countries in the 2000s. It is also clear from our results for some of the datasets that the sex of the child is ranking last.

We also extracted survival curves from the Deepsurv model to establish whether the survival outcome associated with the four socioeconomic factors has become favourable over-time.

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

Figures 6 shows survival curves of the survival outcome (under-five survival time) associated to the four socioeconomic factors extracted from the deep learning survival model for the test datasets obtained from the datasets of the countries representing all the four sub-regions considered in this study. The survival curves show an improvement in the survival probabilities associated to the four socioeconomic factors for the children under the age of five in the country's 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 sub-region had a survival curve for the children under the age of five in the year 2014 above that of the year 2008. And lastly for Chad in the central sub-region, the survival curve for the year 2014 is above that of 2004.

This is an indicator that there is improvement in the survival outcome associated to the four socioeconomic factors in these countries' over-time especially after five and above years of launching the millennium development goals.

All the countries considered for analysis in the different sub-regions had a median survival time associated to the four socioeconomic factors for the children in the test dataset of above five years; however, we noticed that this improvement has been gradual. For example, a country like Uganda in the East African sub-region 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, we noticed that for the year 2011, the survival curves for the children under the age of one year is above that of the children below the same age in 2006. However, the survival curve for children above one year in 2011 compared to those above one 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 outcome associated to the four socioeconomic factors over time.

This is an indicator 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 model 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 below.

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

Figure 7 shows that the mean values of the concordance index 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 to the random survival forest model.

The performance of this model on datasets of a country from each sub-region has no clear trend but what is obvious is that these four socioeconomic factors are still predictive in determining U5MR in sub-Saharan Africa. Infact in some of some 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 to U5MR and therefore public health policies to achieve SDG3 should be designed to target inequalities based on these factors that exist within each country in the sub-regions.

Discussion

The study reveals that the four social economic factors, wealth index (household wealth) and mother's education level are the top contributors of mortality in the countries considered in this study over a period of ten years. 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), Ghana (2008 and 2014). Place of residence ranked first in datasets like Chad (2004 and 2014).

With a mean concordance 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 within the region. Survival curves of the survival outcome associated to the four social economic 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 dataset in the study. For the country like Zimbabwe which is a representative of the Southern African sub-region, 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 to the earlier years in the four countries selected to represent the different sub-regions.

The main strength of this study is that we used machine learning methods which when compared to classical statistical models are very flexible have fewer assumption. 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 overtime, which can 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 that died before the survey which creates respondent bias.

Our results on the most influential factors associated to U5MR agree with studies other studies.[2-3,25,52-54] Ezeh et al.,[54] found out 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- and middle-income countries ranked mother's education and household wealth index among the most critical predictors of U5MR. The same study also 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 country overtime and presented how the survival outcome associated to the economic factors has improved overtime.

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 policy makers 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 overtime. It will be interesting to follow the ranking of all the factors that are sociated to U5MR in the region. It would also be interesting to see how the survival outcome is improving overtime after considering all the other factors that determine U5MR in the region. The study also excluded some of the datasets within the countries chosen for analysis, most among them were 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 millennium development goals to improve the survival outcome of children under the age of five 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 Demographic health surveys and the use of machine learning techniques, we can uncover a lot of policy signals. If used well, this information can guide policymakers on what policies to implement and what sectors to target to achieve the sustainable development goals. For example, our study has looked at how ranks of importance, the survival outcome, and the predictive nature of four social economic determinants of U5MR has evolved using two machine learning techniques. The results have uncovered interesting results that can be used to inform policy on what sectors to target to achieve SDG3. The study has revealed that most of the policies should target reducing poverty levels and aim at increasing literacy level of the girl child in the region. The study has also revealed that the past interventions aimed at targeting these four social economic factors are starting to pay-off. This is because over-time the survival outcome associated to these factors has become more and more favourable.

The DeepSurv model has higher predictive performance of with mean concordance index values (between 67% to 80%), above 50%, indicating that these factors are still highly associated to U5MR. Therefore, this study is advocating for reviewing the success of these policies using machine learning methods to know where to put much effort along the implementation process of these policies 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.

Availability of data

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 program.

Author's contribution

JBN and HM conceptualised the study, JBN conducted the data extraction, JN and RM trained the models on the datasets and wrote the first draft of the manuscript. HM edited and proofread the document.

Competing risks

None declared.

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. 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.

Patient consent for publication

Not required.

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).

Acknowledgements

The first and second author 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 Program 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.

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(1):1003.

2. Tabutin D, Masquelier B, Grieve M, et al. Mortality Inequalities and Trends in Low- and Middle-Income Countries, 1990–2015. Population, English edition. 2017;72(2):221 – 295.

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(1):760.

4. Mosley WH, Chen LC. An Analytical Framework for the Study of Child Survival in Developing Countries. Population and Development Review. 1984; 10:25–45.

5. Satagopan JM, Ben-Porat L, Berwick M, et al. A note on competing risks in survival data analysis. British Journal of Cancer. 2004;91(7):1229–1235.

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 Nutrition. 2017;3(1):18.

7. Sahu D, Nair S, Singh L, et al. Levels, trends & predictors of infant & child mortality among Scheduled Tribes in rural India. The Indian journal of medical research. 2015;141(5):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 Pacific journal of clinical nutrition. 2012;21(4):568.

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(1).

10. Kanmiki EW, Bawah AA, Agorinya I, et al. Socio-economic and demographic determinants of under-five mortality in rural northern Ghana. BMC international health and human rights. 2014;14(1):24.

11. Ayele DG, Zewotir TT, Mwambi H. Survival analysis of under-five mortality using Cox and frailty models in Ethiopia. Journal of Health, Population and Nutrition. 2017;36(1):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 and childbirth. 2012;12(1):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(8).

14. Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. Biometrika. 1994;81(3):515–526.

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 Medical Research Methodology. 2017;17(1):115.

16. Faraggi D, Simon R. A neural network model for survival data. Statistics in Medicine. 1995;14(1):73–82.

17. Ishwaran H, Kogalur UB, Blackstone EH, et al. Random survival forests. Annals of Applied Statistics. 2008;2(3):841–860.

18. Yousefi S, Amrollahi F, Amgad M, et al. Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. Scientific Reports. 2017;7(1):11707.

19. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–444.

20. Luck M, Sylvain T, Cardinal H, et al. Deep Learning for Patient-Specific Kidney Graft Survival Analysis. arXiv:170510245 [cs, stat]. 2017.

21. Katzman JL, Shaham U, Cloninger A, et al. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Medical Research Methodology. 2018;18(1):24.

22. Sargent DJ. Comparison of artificial neural networks with other statistical approaches: results from medical data sets. Cancer. 2001;91(8):1636–1642.

23. Xiang A, Lapuerta P, Ryutov A, et al. Comparison of the performance of neural network methods and Cox regression for censored survival data. Computational Statistics & Data Analysis. 2000;34(2):243–257.

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 Research and Treatment. 1997;44(2):167–178.

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 Aug 1;10(8): 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 journal of community medicine: official publication of Indian Association of Preventive & Social Medicine. 2016;41(1):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 research notes. 2017;10(1):459.

28. Breiman L, Friedman J, Stone CJ, et al. Classification and regression trees; 1984.

29. Morgan JN, Sonquist JA. Problems in the analysis of survey data, and a proposal. Journal of the American statistical association. 1963;58(302):415–434.

30. Gordon L, Olshen R. Tree-structured survival analysis. Cancer treatment reports. 1985;69(10):1065–1069.

31. Bou-Hamad I, Larocque D, Ben-Ameur H, et al. A review of survival trees. Statistics Surveys. 2011; 5:44–71.

32. Breiman L. Random forests. Machine learning. 2001;45(1):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(3):651–674.

35. Wright MN, Dankowski T, Ziegler A. Unbiased split variable selection for random survival forests using maximally selected rank statistics. Statistics in medicine. 2017;36(8):1272–1284.

36. Wright MN, Ziegler A. ranger: A fast implementation of random forests for high dimensional data in C++ and R. Journal of Statistical Software. 2017;77(i01).

37. R Core Team. R: A Language and Environment for Statistical Computing. https://www.R-project.org/. R Foundation for Statistical Computing. 2013.

38. Ishwaran H, Kogalur UB, Kogalur MU, Suggests XM. Package 'randomSurvivalForest'.. 2013.

39. Cox DR. Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological). 1972;34(2):187–202.

40. Fotso, S. "PySurvival: open-source package for survival analysis modeling." 2019.

41. Harrell Jr FE, Lee KL, Califf RM, et al. Regression modelling strategies for improved prognostic prediction. Statistics in medicine. 1984;3(2):143–152.

42. G¨onen M, Heller G. Concordance probability and discriminatory power in proportional hazards regression. Biometrika. 2005;92(4):965–970.

43. Pencina MJ, D'Agostino RB. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. Statistics in medicine. 2004;23(13):2109–2123.

44. Santos MY, e Sa´ JO, Andrade C, et al. A big data system supporting bosch braga industry 4.0 strategy. International Journal of Information Management. 2017;37(6):750–760.

45. Schwarz DF, K¨onig IR, Ziegler A. On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data. Bioinformatics. 2010;26(14):1752–1758.

46. Jones Z, Linder F. Exploratory data analysis using random forests. In: Prepared for the 73rd annual MPSA conference; 2015.

47. Ishwaran H. Variable importance in binary regression trees and forests. Electronic Journal of Statistics. 2007; 1:519–537.

48. Ishwaran H, Kogalur UB, Gorodeski EZ, et al. High-dimensional variable selection for survival data. Journal of the American Statistical Association. 2010;105(489):205–217.

49. Strobl C, Boulesteix A, Zeileis A, et al. Bias in random forest variable importance measures: Illustrations, sources, and a solution. BMC bioinformatics. 2007;8(1):25.

50. Wright MN, Ziegler A, K¨onig IR. Do little interactions get lost in dark random forests? BMC bioinformatics. 2016;17(1):145.

51. Strobl C, Boulesteix A, Kneib T, et al. Conditional variable importance for random forests. BMC Bioinformatics. 2008;9(1):307.

52. Rutstein S, Winter R, Staveteig S, et al. Urban Child Poverty, Health, and Survival in Low-and Middle-income Countries. In: PAA 2017 Annual Meeting; 2017.

53. Kunst AE, Mackenbach JP. The size of mortality differences associated with educational level in nine industrialized countries. American journal of public health. 1994;84(6):932–937.

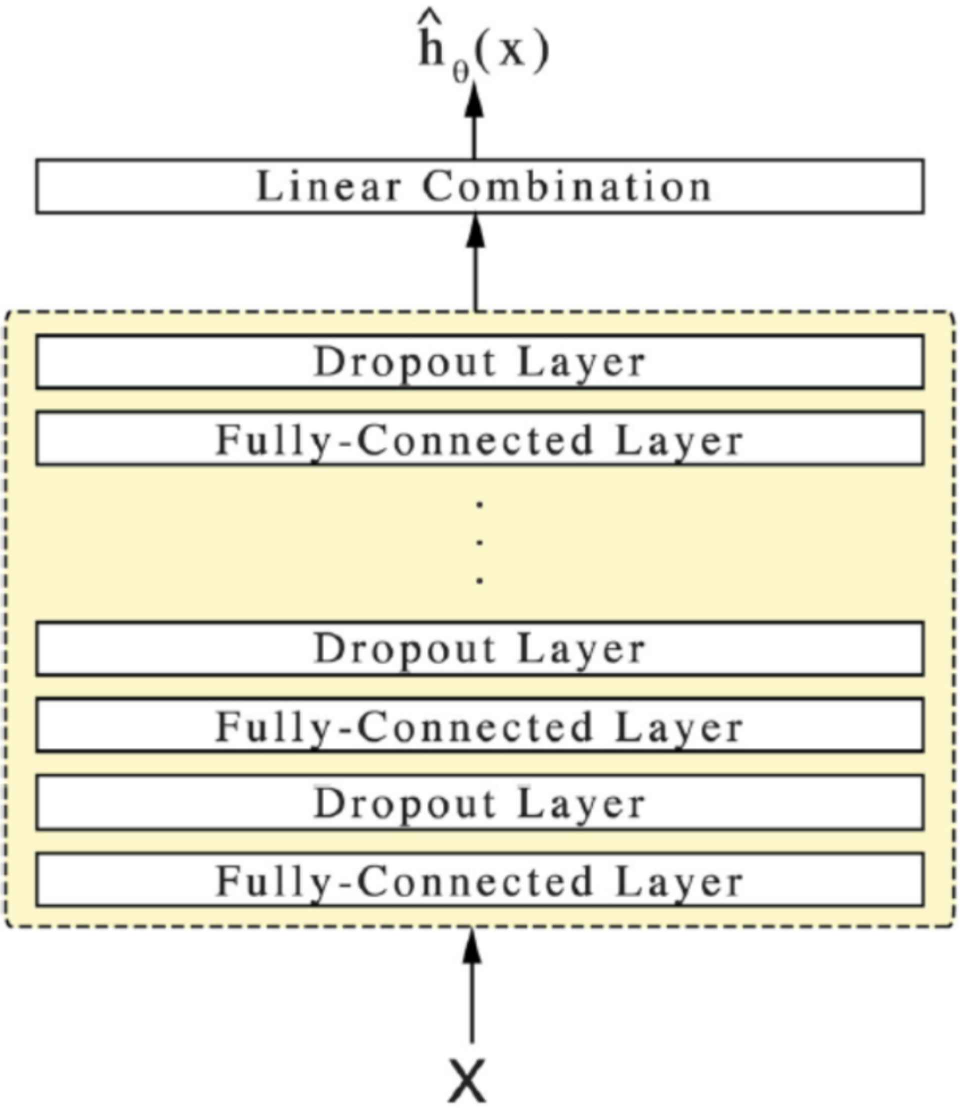
54. Ezeh OK, Agho KE, Dibley MJ, et al. Risk factors for postneonatal, infant, child, and under-5 mortality in Nigeria: a pooled cross-sectional analysis. BMJ open. 2015;(5)3: e006779

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. Academic emergency medicine, 2016(23)3: p. 269-278.

56. Panesar SS, D'Souza RN, Yeh FC, et al. Machine learning versus logistic regression methods for 2-year mortality prognostication in a small, heterogeneous glioma database. World neurosurgery*:* 2019(2): 100012.

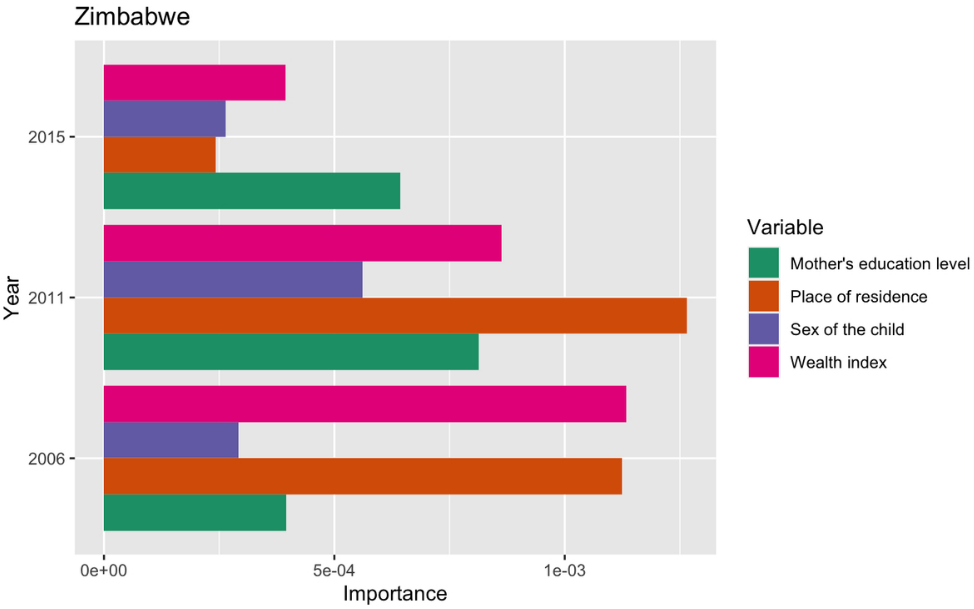
57. Kimani-Murage EW, Fotso JC, Egondi T, et al. Trends in childhood mortality in Kenya: the urban advantage has seemingly been wiped out. Health & place. 2014; 29:95–103.

58. Sousa A, Hill K, Dal Poz MR. Sub-national assessment of inequality trends in neonatal and child mortality in Brazil. International journal for equity in health. 2010;9(1):21.



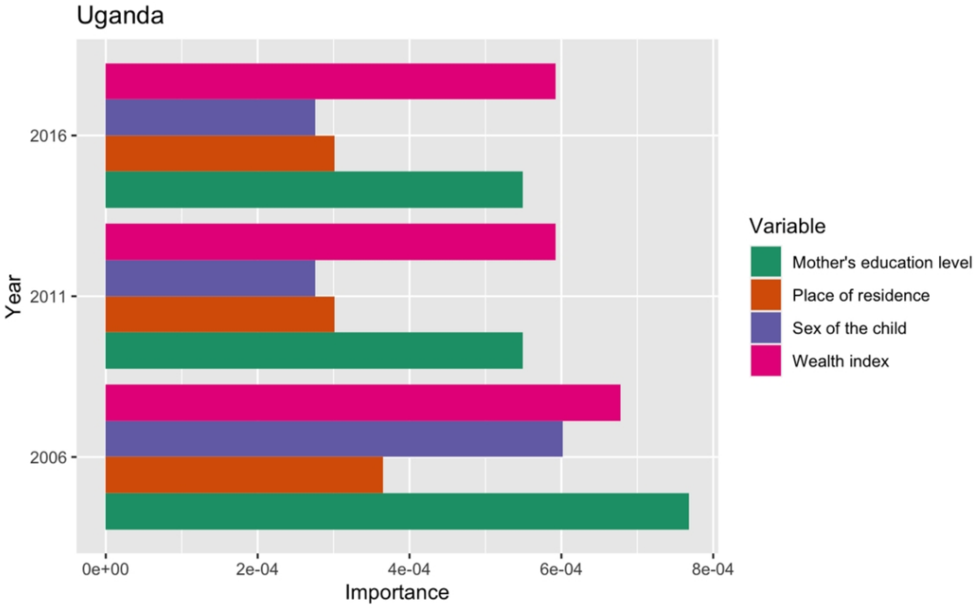
. DeepSurv architecture Katzman et al. [18]

75x84mm (300 x 300 DPI)



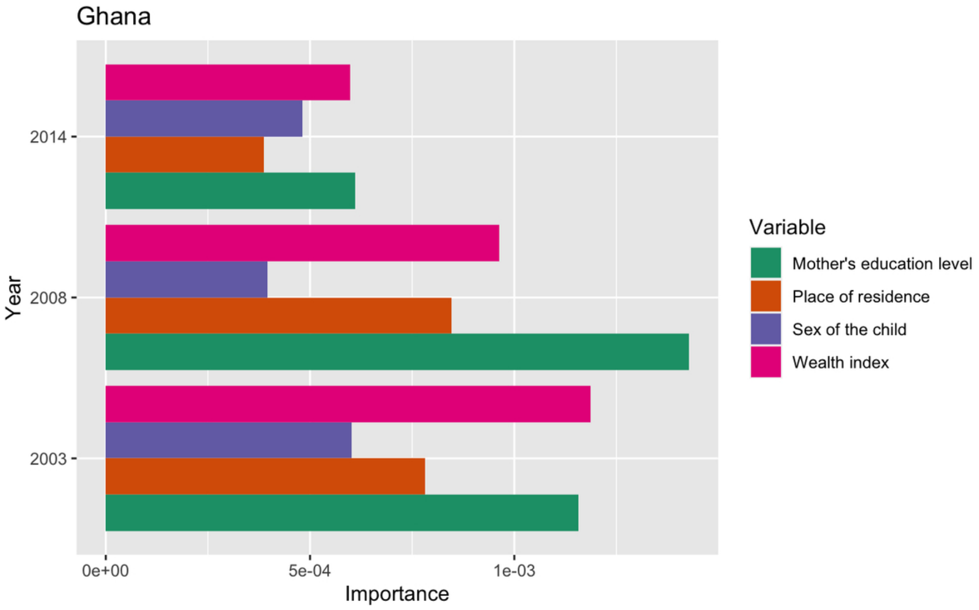
Ranks of importance for the four socioeconomic factors in predicting U5MR in Zimbabwe over the period of 9 years

101x62mm (300 x 300 DPI)



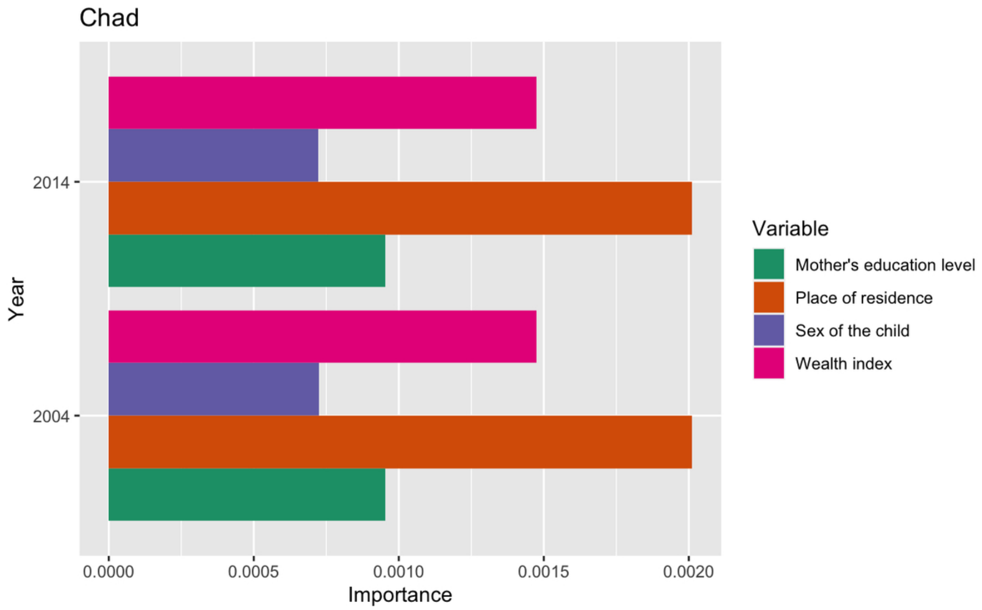
Ranks of importance for the four social economic factors in predicting U5MR in Uganda over the period of 10 years

101x62mm (300 x 300 DPI)



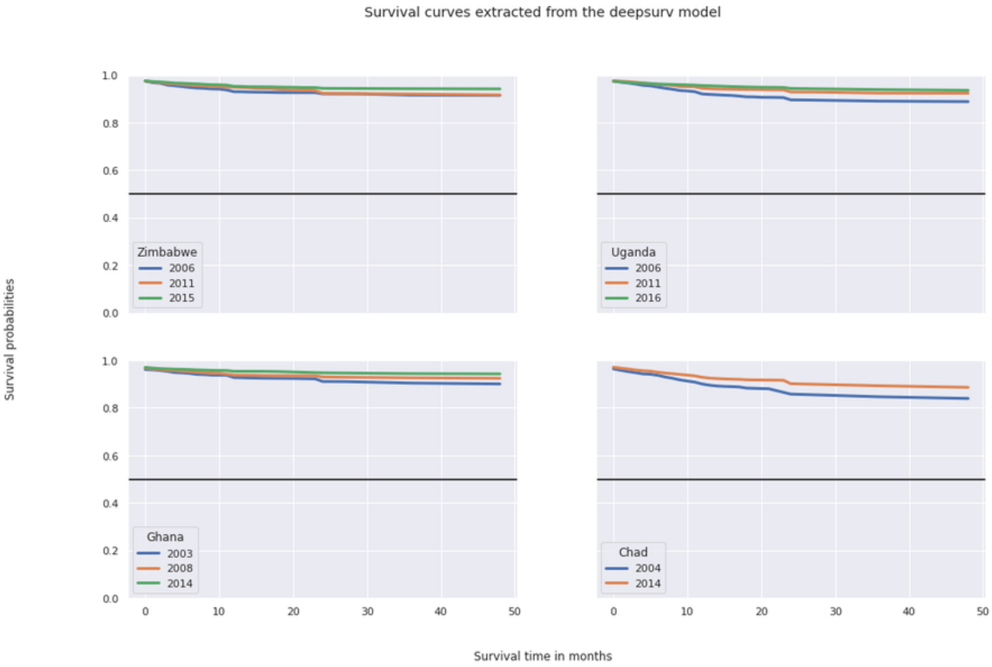
Ranks of importance for the four social economic factors in predicting U5MR in Ghana over the period of 10 years

101x62mm (300 x 300 DPI)



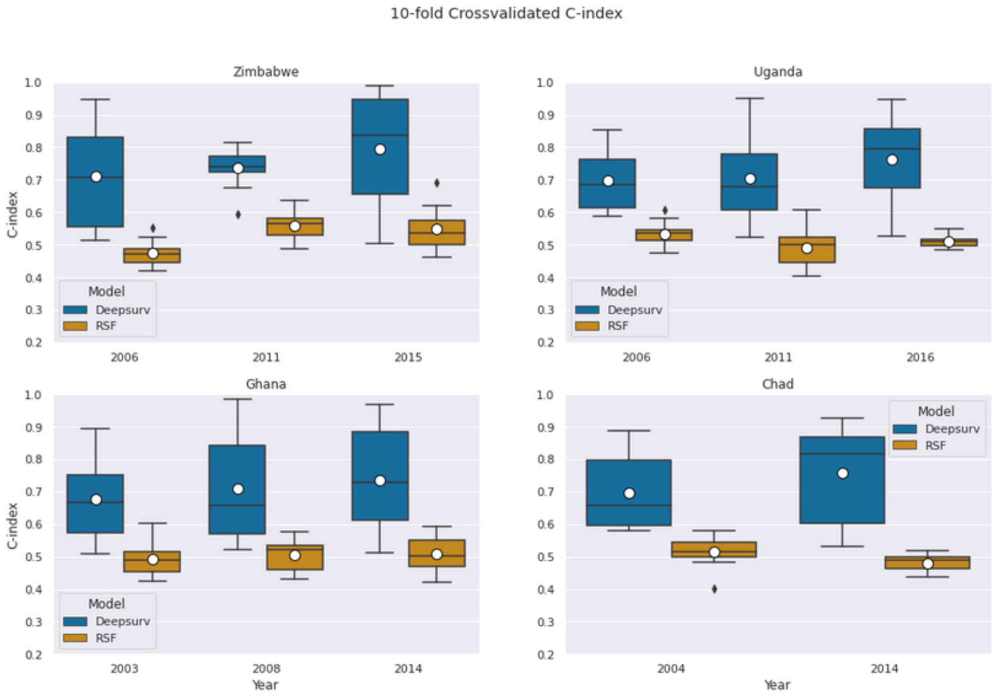
Ranks of importance for the four social economic factors in predicting U5MR in Chad over the period of 10 years

101x62mm (300 x 300 DPI)



Survival probabilities for the children in the test dataset for Zimbabwe, Uganda, Ghana, and Chad obtained from the deepsurv model

75x50mm (300 x 300 DPI)



Comparison of predictive performance of the deep survival neural network and the random survival forest models on all the datasets considered in this study

68x48mm (300 x 300 DPI)

STROBE Statement—checklist of items that should be included in reports of observational studies

	Item No.	Recommendation	Page No.	Relevant text from manuscript
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract	1	Abstract
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found	1	Abstract
Introduction				
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported	2 -3	Introduction
Objectives	3	State specific objectives, including any prespecified hypotheses	3	Introduction
Methods				
Study design	4	Present key elements of study design early in the paper	3	Data
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	3-6	Data
Participants	6	*(a)* *Cross-sectional study*—Give the eligibility criteria, and the sources and methods of selection of participants	3-6	Data
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable	6	Data and Data pre- processing
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	*3-6*	Introduction and Data
Bias	9	Describe any efforts to address potential sources of bias	3- 6	Data and Data pre- processing
Study size	10	Explain how the study size was arrived at	1	Abstract

Continued on next page

Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why	5 -6	Data pre-processing
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding	7-11	Models
		(*b*) Describe any methods used to examine subgroups and interactions		N/A
		(*c*) Explain how missing data were addressed	5-6	Data pre-processing
		(*d*) *Cross-sectional study*—If applicable, describe analytical methods taking account of sampling strategy		N/A
		(*e*) Describe any sensitivity analyses		N/A
Results				
Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed	3-6	Data
		(b) Give reasons for non-participation at each stage	3-6	Data
		(c) Consider use of a flow diagram		N/A
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders	Table 2 and 3	Page 5 and 6
		(b) Indicate number of participants with missing data for each variable of interest	N/A	
		(c) *Cohort study*—Summarise follow-up time (eg, average and total amount)		
Outcome data	15*	*Cohort study*—Report numbers of outcome events or summary measures over time		
		Case-control study—Report numbers in each exposure category, or summary measures of exposure		
		Cross-sectional study—Report numbers of outcome events or summary measures	3- 6	Data and Data pre- processing
Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included	12-17	Results
		(*b*) Report category boundaries when continuous variables were categorized	N/A	
		(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	N/A	

Continued on next page

Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses	N/A	
Discussion				
Key results	18	Summarise key results with reference to study objectives	17-18	Discussion
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	17-18	Discussion
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	18-19	Discussion
Generalisability	21	Discuss the generalisability (external validity) of the study results	19	Conclusion
Other information				
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	20	Acknowledgement

*Give information separately for cases and controls in case-control studies and, if applicable, for exposed and unexposed groups in cohort and cross-sectional studies.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article

Standards for Reporting Qualitative Research (SRQR)*

http://www.equator-network.org/reporting-guidelines/srqr/

Page/line no(s).

Title and abstract

Title - Concise description of the nature and topic of the study Identifying the study as qualitative or indicating the approach (e.g., ethnography, grounded theory) or data collection methods (e.g., interview, focus group) is recommended	PAGE 1
Abstract - Summary of key elements of the study using the abstract format of the intended publication; typically includes background, purpose, methods, results, and conclusions	PAGE 1

Introduction

Problem formulation - Description and significance of the problem/phenomenon studied; review of relevant theory and empirical work; problem statement	PAGE 2
Purpose or research question - Purpose of the study and specific objectives or questions	PAGE 3

Methods

Qualitative approach and research paradigm - Qualitative approach (e.g., ethnography, grounded theory, case study, phenomenology, narrative research) and guiding theory if appropriate; identifying the research paradigm (e.g., postpositivist, constructivist/ interpretivist) is also recommended; rationale**	PAGE 7 TO 12
Researcher characteristics and reflexivity - Researchers' characteristics that may influence the research, including personal attributes, qualifications/experience, relationship with participants, assumptions, and/or presuppositions; potential or actual interaction between researchers' characteristics and the research questions, approach, methods, results, and/or transferability	PAGE 11 TO 17
Context - Setting/site and salient contextual factors; rationale**	
Sampling strategy - How and why research participants, documents, or events were selected; criteria for deciding when no further sampling was necessary (e.g., sampling saturation); rationale**	PAGE 4 TO 6
Ethical issues pertaining to human subjects - Documentation of approval by an appropriate ethics review board and participant consent, or explanation for lack thereof; other confidentiality and data security issues	PAGE 20
Data collection methods - Types of data collected; details of data collection procedures including (as appropriate) start and stop dates of data collection and analysis, iterative process, triangulation of sources/methods, and modification of procedures in response to evolving study findings; rationale**	PAGE 3 TO 6 AND PAGE 20

Data collection instruments and technologies - Description of instruments (e.g., interview guides, questionnaires) and devices (e.g., audio recorders) used for data collection; if/how the instrument(s) changed over the course of the study	N/A
Units of study - Number and relevant characteristics of participants, documents, or events included in the study; level of participation (could be reported in results)	PAGE 20
Data processing - Methods for processing data prior to and during analysis, including transcription, data entry, data management and security, verification of data integrity, data coding, and anonymization/de-identification of excerpts	N/A
Data analysis - Process by which inferences, themes, etc., were identified and developed, including the researchers involved in data analysis; usually references a specific paradigm or approach; rationale**	PAGE 12 TO 17
Techniques to enhance trustworthiness - Techniques to enhance trustworthiness and credibility of data analysis (e.g., member checking, audit trail, triangulation); rationale**	PAGE 20

Results/findings

Synthesis and interpretation - Main findings (e.g., interpretations, inferences, and themes); might include development of a theory or model, or integration with prior research or theory	PAGE 18 TO 19
Links to empirical data - Evidence (e.g., quotes, field notes, text excerpts, photographs) to substantiate analytic findings	PAGE 18 TO 19

Discussion

Integration with prior work, implications, transferability, and contribution(s) to the field - Short summary of main findings; explanation of how findings and conclusions connect to, support, elaborate on, or challenge conclusions of earlier scholarship; discussion of scope of application/generalizability; identification of unique contribution(s) to scholarship in a discipline or field	PAGE 18 TO 19
Limitations - Trustworthiness and limitations of findings	PAGE 1 TO 2

Other

Conflicts of interest - Potential sources of influence or perceived influence on study conduct and conclusions; how these were managed	PAGE 20
Funding - Sources of funding and other support; role of funders in data collection, interpretation, and reporting	PAGE 20

*The authors created the SRQR by searching the literature to identify guidelines, reporting standards, and critical appraisal criteria for qualitative research; reviewing the reference lists of retrieved sources; and contacting experts to gain feedback. The SRQR aims to improve the transparency of all aspects of qualitative research by providing clear standards for reporting qualitative research.

**The rationale should briefly discuss the justification for choosing that theory, approach, method, or technique rather than other options available, the assumptions and limitations implicit in those choices, and how those choices influence study conclusions and transferability. As appropriate, the rationale for several items might be discussed together.

Reference:

O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. **Standards for reporting qualitative research: a synthesis of recommendations.** *Academic Medicine*, Vol. 89, No. 9 / Sept 2014 DOI: 10.1097/ACM.0000000000000388

COREQ (COnsolidated criteria for REporting Qualitative research) Checklist

A checklist of items that should be included in reports of qualitative research. You must report the page number in your manuscript where you consider each of the items listed in this checklist. If you have not included this information, either revise your manuscript accordingly before submitting or note N/A.

Topic	Item No.	Guide Questions/Description	Reported on Page No.
Domain 1: Research team and reflexivity			
Personal characteristics			
Interviewer/facilitator	1	Which author/s conducted the interview or focus group?	DHS- 3 TO 4
Credentials	2	What were the researcher's credentials? E.g. PhD, MD	DHS- 3 TO 4
Occupation	3	What was their occupation at the time of the study?	N/A
Gender	4	Was the researcher male or female?	BOTH MALE AN
Experience and training	5	What experience or training did the researcher have?	DHS- 3 TO 4
Relationship with participants			
Relationship established	6	Was a relationship established prior to study commencement?	N/A
Participant knowledge of the interviewer	7	What did the participants know about the researcher? e.g. personal goals, reasons for doing the research	DHS- 3 TO 4
Interviewer characteristics	8	What characteristics were reported about the inter viewer/facilitator? e.g. Bias, assumptions, reasons and interests in the research topic	DHS- 3 TO 4
Domain 2: Study design			
Theoretical framework			
Methodological orientation and Theory	9	What methodological orientation was stated to underpin the study? e.g. grounded theory, discourse analysis, ethnography, phenomenology, content analysis	PAGE 2 TO 3
Participant selection			
Sampling	10	How were participants selected? e.g. purposive, convenience, consecutive, snowball	DHS- 3 TO 4
Method of approach	11	How were participants approached? e.g. face-to-face, telephone, mail, email	DHS- 3 TO 4
Sample size	12	How many participants were in the study?	TABLE 2 AND 3
Non-participation	13	How many people refused to participate or dropped out? Reasons?	DHS- 3 TO 4
Setting			
Setting of data collection	14	Where was the data collected? e.g. home, clinic, workplace	DHS- 3 TO 4
Presence of non-participants	15	Was anyone else present besides the participants and researchers?	DHS- 3 TO 4
Description of sample	16	What are the important characteristics of the sample? e.g. demographic data, date	PAGE 3 TO 6
Data collection			
Interview guide	17	Were questions, prompts, guides provided by the authors? Was it pilot tested?	DHS- 3 TO 4
Repeat interviews	18	Were repeat inter views carried out? If yes, how many?	N/A
Audio/visual recording	19	Did the research use audio or visual recording to collect the data?	NO
Field notes	20	Were field notes made during and/or after the inter view or focus group?	DHS- 3 TO 4
Duration	21	What was the duration of the inter views or focus group?	N/A
Data saturation	22	Was data saturation discussed?	N/A
Transcripts returned	23	Were transcripts returned to participants for comment and/or	N/A

Topic	Item No.	Guide Questions/Description	Reported on Page No.
		correction?	
Domain 3: analysis and findings			
Data analysis			
Number of data coders	24	How many data coders coded the data?	ONE
Description of the coding tree	25	Did authors provide a description of the coding tree?	NO
Derivation of themes	26	Were themes identified in advance or derived from the data?	DERIVED FROM
Software	27	What software, if applicable, was used to manage the data?	R AND PYTHON
Participant checking	28	Did participants provide feedback on the findings?	NO
Reporting			
Quotations presented	29	Were participant quotations presented to illustrate the themes/findings? Was each quotation identified? e.g. participant number	NO
Data and findings consistent	30	Was there consistency between the data presented and the findings?	YES
Clarity of major themes	31	Were major themes clearly presented in the findings?	YES
Clarity of minor themes	32	Is there a description of diverse cases or discussion of minor themes?	YES-PAGE 18

Developed from: Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. *International Journal for Quality in Health Care*. 2007. Volume 19, Number 6: pp. 349 – 357

Once you have completed this checklist, please save a copy and upload it as part of your submission. DO NOT include this checklist as part of the main manuscript document. It must be uploaded as a separate file.

BMJ Open

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

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2021-049786.R2
Article Type:	Original research
Date Submitted by the Author:	22-Dec-2021
Complete List of Authors:	Nasejje, Justine B; University of the Witwatersrand, Statistics and Actuarial Science Mbuvha, Rendani; University of the Witwatersrand, Statistics and Actuarial Science Mwambi, Henry; University of Kwazulu-Natal, School of Mathematics, Statistics and Computer Science
<b>Primary Subject Heading</b>:	Public health
Secondary Subject Heading:	Health policy, Global health, Public health
Keywords:	Health policy < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, PUBLIC HEALTH, Community child health < PAEDIATRICS

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

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

1 School of Statistics and Actuarial Science, University of Witwatersrand, Jan Smuts Avenue, Johannesburg, Gauteng, South Africa

2 School of Statistics, Mathematics and Computer Science, University of KwaZulu-Natal, King Edward Avenue, Pietermaritzburg, South Africa

* Corresponding author E-mail: 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-five 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 sub-regions of sub-Saharan Africa.

Participants: Each country has multiple DHS datasets and a total of eleven 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% to 80%), above 50% compared to the RSF model.

Conclusions: The study reveals that children under the age of five 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 (SDG3).

Strengths and Limitations of the Study

- The study used machine learning methods which when compared to 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-five 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 five is a global indicator of societal and national development; it serves as a key marker of health equity and access.[1] The fourth Millennium Development

Goal (MDG4), which centred at reducing under-five mortality by two-thirds in the period between 1990 and 2015, now appears in the third Sustainable Development Goal (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 sub-groups 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, semi-parametric 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 socio-demographic 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 semi-parametric 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 to 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- 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-five mortality in the Sub-Saharan region.[12-25] With the launch of the millennium development goals in the year 2000, we saw the convergence of the development agendas of United Nations Development Programme (UNDP); United Nations Environment Programme (UNEP); World Health Organisation (WHO); United Nations Children's Fund (UNICEF); United Nations Educational, Scientific and Cultural Organization (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 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 random survival forest 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 five 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 random survival forest 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 sub-regions (Southern, Central, Eastern and Western Africa) of sub-Saharan Africa. DHS is funded by USAID, UNFPA, UNICEF, Irish Aid and the government of the United Kingdom 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 program website. The datasets are available on request from the DHS program. The outcome variable is under-five 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 five 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 education, wealth index of the household, and sex of the child. The four countries and the demographic health survey datasets selected from each sub-region are shown in Table 1 below.

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

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

Data pre-processing

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 five 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 1990's and the early 2000's, wealth index was not a recorded feature. These datasets were excluded in our final analysis to allow meaningful comparisons. Table 2 and Table 3 give the counts of the number of children under the age of five for each of the feature category in all the datasets considered for analysis.

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 five by wealth index

	Wealth index					Total
	Poorest	Poorer	Middle	Richer	Richest	
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 85,688.

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 random survival forests, 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] 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 random survival forest, and the deep survival neural network model (DeepSurv). [17, 21]

Random survival forests

Random survival forests (RSF) 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 random survival forest model by Breiman et al.[28]

Survival trees

The regression tree algorithm for right censored data, is an extension of the CART algorithm by Breiman et al.,[28]. Algorithm 1 below is the general algorithm for survival trees.[29-31]

```
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 ($\alpha$ and $\beta$) 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 i.e., has no less than $d_0 > 0$ unique observed events.
```

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. Algorithm 2 for building an RSF model as presented by Ishwaran et al.,[17] is given below as follows:

```
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.
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.
```

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 random survival forest 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 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.

Cox proportional hazards model

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

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

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 $\hat{h}_\beta(X) = \hat{\beta}' \cdot X$. The estimates $(\hat{\beta})$ 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:

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

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:

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

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:

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

Where $\gamma > 0$ and $\lambda > 0$ 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 $\hat{h}_\beta(X) = \hat{\beta}' \cdot X$ in equation 2 by the output of $\hat{h}_\theta(X) = \exp\left(g(X,\theta)\right)$ of the neural network. The proportional hazards model becomes

$$h_\theta(X_i) = \exp\left(g(X_i,\theta)\right). \qquad (4)$$

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:

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

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

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

where $n_{\delta 1}$ 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:

$$l(\theta) = -\frac{1}{n_{\delta_1}} \sum_{i:\delta_i = 1} \left(\hat{h}_\theta(X_i) - \log \sum_{j \in \mathfrak{R}(t_i)} \exp\left(\hat{h}_\theta(X_j)\right)\right) + \alpha \parallel \theta \parallel_2^2 \quad (7)$$

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 below shows its architecture. It is a deep feed-forward neural network implemented as:

Fig 1. DeepSurv architecture Katzman et al.,[21].

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 dropout 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 $\hat{Y}$, the C-index is given by:

$$C = \frac{\sum_{i:\delta_i = 1} \sum_{y_i < y_j} I\left(\hat{Y}_i < \hat{Y}_j\right)}{Number\ of\ comparable\ pairs} \qquad (8)$$

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] Over-fitting 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 random survival forests model has a measure which estimates the importance of each covariate. It is generally referred to as the variable importance measure (VIMP).[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 ($\hat{e}$), calculated by using 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, ($\hat{e}_i$) using OOB observations. Compare $\hat{e}_i$, with $\hat{e}$ of the original classification for covariate, X_i. Calculate, argmax $\{0; \hat{e}_i - \hat{e}\}$. 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 below. The ranks of feature importance presented here are for datasets from each country that was selected from each sub-region.

Fig 2. Ranks of importance for the four socioeconomic factors in predicting U5MR in Zimbabwe over a period of 9 years.

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.

Fig 3. Ranks of importance for the four social economic factors in predicting U5MR in Ghana over a10 year period.

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

Fig 4. Ranks of importance for the four social economic factors in predicting U5MR in Uganda over a period of 10 years.

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.

Fig 5. Ranks of importance for the four social economic factors in predicting U5MR in Chad over the period of 10 years.

Figures 2-5 show that mother's education is ranked first in five out of the eleven datasets, and wealth index ranked first in three out of the eleven datasets, but second in eight out of the eleven 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 two out of the eleven datasets, and second in one of the eleven datasets, placing it among the top three predictors of under-five 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- 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.

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

Figures 6 shows survival curves of the survival outcome (under-five 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 sub-regions 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 five in the countries over-time. Zimbabwe, in the southern African sub-region, 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 sub-region, had a survival curve for the children under the age of five in the year 2014 above that of the year 2008. And lastly, for Chad, in the central sub-region, 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 five or more years after the launch of the millennium development goals.

The countries considered for analysis in the different sub-regions had a median survival time associated to the four socioeconomic factors for the children in the test dataset of above five years; however, we noticed that this improvement has been gradual. For example, a country like Uganda from the East African sub-region 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 one year is above that of the children below the same age in 2006. However, the survival curve for children above one year in 2011 compared to those above one 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 concordance 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 below.

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

Figure 7 shows that the mean values of the cross-validated concordance 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 to the random survival forest model.

The performance of this model on datasets of a country from each sub-region 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 concordance 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 sub-region, 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 to the earlier years in the four countries selected from the different sub-regions.

The main strength of this study is that we used machine learning methods which, when compared to 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- 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 twenty-five years and therefore policy makers 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 millennium development goals to improve the survival outcome of children under the age of five 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 demographic health surveys and the use of machine learning techniques, we can uncover a lot of policy signals. If used well, this information can guide policymakers on what policies to implement and what sectors to target to achieve the sustainable development goals. 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 concordance 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 programs 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.

Availability of data

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 program.

Authors 'contribution

JBN and HM conceptualised the study, JBN conducted the data extraction, JN and RM trained the models on the datasets and wrote the first draft of the manuscript. HM edited and proofread the document.

Competing risks

None declared.

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. 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.

Patient consent for publication

Not required.

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).

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 Program 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.

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(1):1003.
2. Tabutin D, Masquelier B, Grieve M, et al. Mortality Inequalities and Trends in Low- and Middle-Income Countries, 1990–2015. Population, English edition. 2017;72(2):221 – 295.
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(1):760.
4. Mosley WH, Chen LC. An Analytical Framework for the Study of Child Survival in Developing Countries. Population and Development Review. 1984; 10:25–45.
5. Satagopan JM, Ben-Porat L, Berwick M, et al. A note on competing risks in survival data analysis. British Journal of Cancer. 2004;91(7):1229–1235.
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 Nutrition. 2017;3(1):18.
7. Sahu D, Nair S, Singh L, et al. Levels, trends & predictors of infant & child mortality among Scheduled Tribes in rural India. The Indian journal of medical research. 2015;141(5):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 Pacific journal of clinical nutrition. 2012;21(4):568.

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(1).
10. Kanmiki EW, Bawah AA, Agorinya I, et al. Socio-economic and demographic determinants of under-five mortality in rural northern Ghana. BMC international health and human rights. 2014;14(1):24.
11. Ayele DG, Zewotir TT, Mwambi H. Survival analysis of under-five mortality using Cox and frailty models in Ethiopia. Journal of Health, Population and Nutrition. 2017;36(1):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 and childbirth. 2012;12(1):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(8).
14. Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. Biometrika. 1994;81(3):515–526.
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 Medical Research Methodology. 2017;17(1):115.
16. Faraggi D, Simon R. A neural network model for survival data. Statistics in Medicine. 1995;14(1):73–82.
17. Ishwaran H, Kogalur UB, Blackstone EH, et al. Random survival forests. Annals of Applied Statistics. 2008;2(3):841–860.
18. Yousefi S, Amrollahi F, Amgad M, et al. Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. Scientific Reports. 2017;7(1):11707.
19. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–444.
20. Luck M, Sylvain T, Cardinal H, et al. Deep Learning for Patient-Specific Kidney Graft Survival Analysis. arXiv:170510245 [cs, stat]. 2017.
21. Katzman JL, Shaham U, Cloninger A, et al. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Medical Research Methodology. 2018;18(1):24.
22. Sargent DJ. Comparison of artificial neural networks with other statistical approaches: results from medical data sets. Cancer. 2001;91(8):1636–1642.
23. Xiang A, Lapuerta P, Ryutov A, et al. Comparison of the performance of neural network methods and Cox regression for censored survival data. Computational Statistics & Data Analysis. 2000;34(2):243–257.
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 Research and Treatment. 1997;44(2):167–178.
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 Aug 1;10(8): 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 journal of community medicine: official publication of Indian Association of Preventive & Social Medicine. 2016;41(1):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 research notes. 2017;10(1):459.
28. Breiman L, Friedman J, Stone CJ, et al. Classification and regression trees; 1984.
29. Morgan JN, Sonquist JA. Problems in the analysis of survey data, and a proposal. Journal of the American statistical association. 1963;58(302):415–434.
30. Gordon L, Olshen R. Tree-structured survival analysis. Cancer treatment reports. 1985;69(10):1065–1069.
31. Bou-Hamad I, Larocque D, Ben-Ameur H, et al. A review of survival trees. Statistics Surveys. 2011; 5:44–71.
32. Breiman L. Random forests. Machine learning. 2001;45(1):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(3):651–674.
35. Wright MN, Dankowski T, Ziegler A. Unbiased split variable selection for random survival forests using maximally selected rank statistics. Statistics in medicine. 2017;36(8):1272–1284.
36. Wright MN, Ziegler A. ranger: A fast implementation of random forests for high dimensional data in C++ and R. Journal of Statistical Software. 2017;77(i01).

37. R Core Team. R: A Language and Environment for Statistical Computing. https://www.R-project.org/. R Foundation for Statistical Computing. 2013.
38. Ishwaran H, Kogalur UB, Kogalur MU, Suggests XM. Package ‘randomSurvivalForest’. 2013.
39. Cox DR. Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological). 1972;34(2):187–202.
40. Fotso, S. "PySurvival: open-source package for survival analysis modeling." 2019.
41. Harrell Jr FE, Lee KL, Califf RM, et al. Regression modelling strategies for improved prognostic prediction. Statistics in medicine. 1984;3(2):143–152.
42. G¨onen M, Heller G. Concordance probability and discriminatory power in proportional hazards regression. Biometrika. 2005;92(4):965–970.
43. Pencina MJ, D’Agostino RB. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. Statistics in medicine. 2004;23(13):2109–2123.
44. Santos MY, e Sa´ JO, Andrade C, et al. A big data system supporting bosch braga industry 4.0 strategy. International Journal of Information Management. 2017;37(6):750–760.
45. Schwarz DF, K¨onig IR, Ziegler A. On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data. Bioinformatics. 2010;26(14):1752–1758.
46. Jones Z, Linder F. Exploratory data analysis using random forests. In: Prepared for the 73rd annual MPSA conference; 2015.
47. Ishwaran H. Variable importance in binary regression trees and forests. Electronic Journal of Statistics. 2007; 1:519–537.
48. Ishwaran H, Kogalur UB, Gorodeski EZ, et al. High-dimensional variable selection for survival data. Journal of the American Statistical Association. 2010;105(489):205–217.
49. Strobl C, Boulesteix A, Zeileis A, et al. Bias in random forest variable importance measures: Illustrations, sources, and a solution. BMC bioinformatics. 2007;8(1):25.
50. Wright MN, Ziegler A, K¨onig IR. Do little interactions get lost in dark random forests? BMC bioinformatics. 2016;17(1):145.
51. Strobl C, Boulesteix A, Kneib T, et al. Conditional variable importance for random forests. BMC Bioinformatics. 2008;9(1):307.
52. Rutstein S, Winter R, Staveteig S, et al. Urban Child Poverty, Health, and Survival in Low-and Middle-income Countries. In: PAA 2017 Annual Meeting; 2017.
53. Kunst AE, Mackenbach JP. The size of mortality differences associated with educational level in nine industrialized countries. American journal of public health. 1994;84(6):932–937.
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)3: 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. Academic emergency medicine, 2016(23)3: p. 269-278.
56. Panesar SS, D'Souza RN, Yeh FC, et al. Machine learning versus logistic regression methods for 2-year mortality prognostication in a small, heterogeneous glioma database. World neurosurgery: 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. International journal for equity in health. 2010;9(1):21.

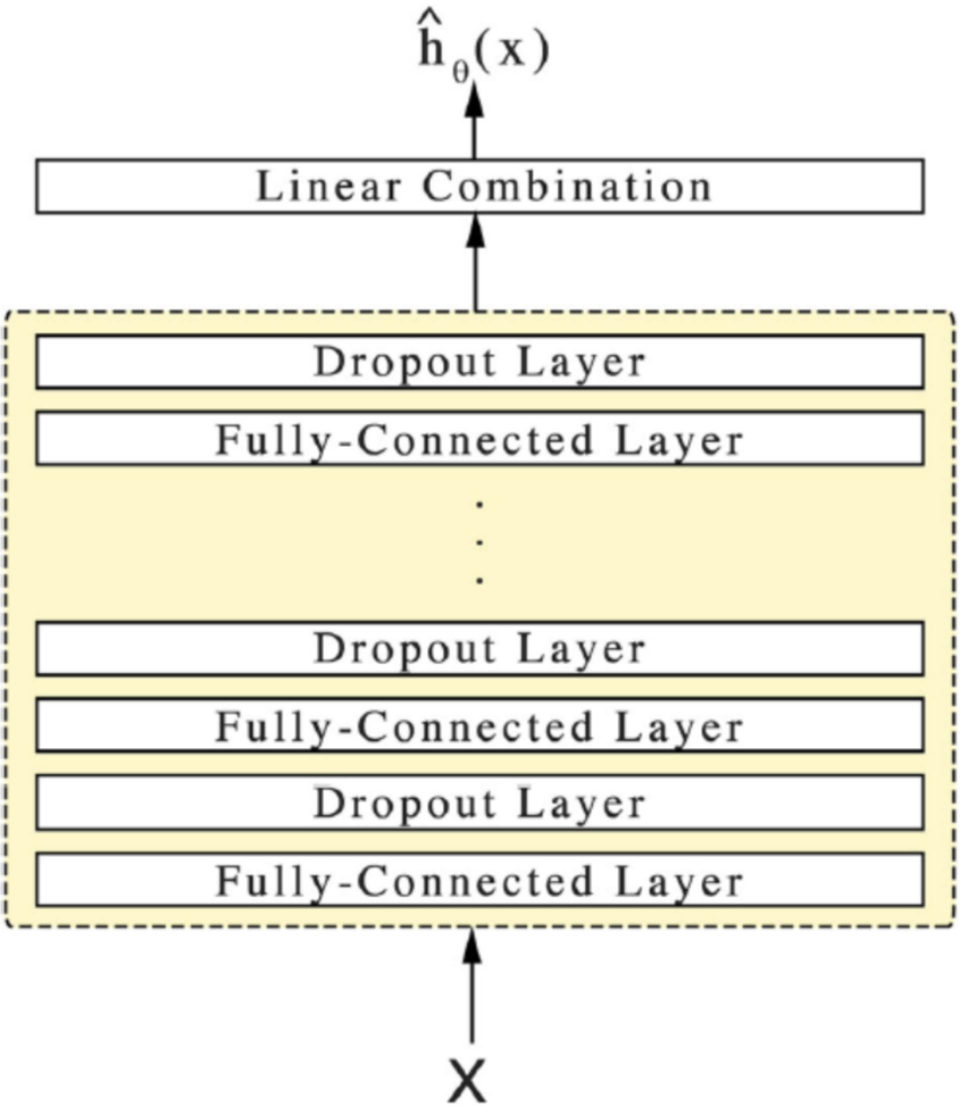
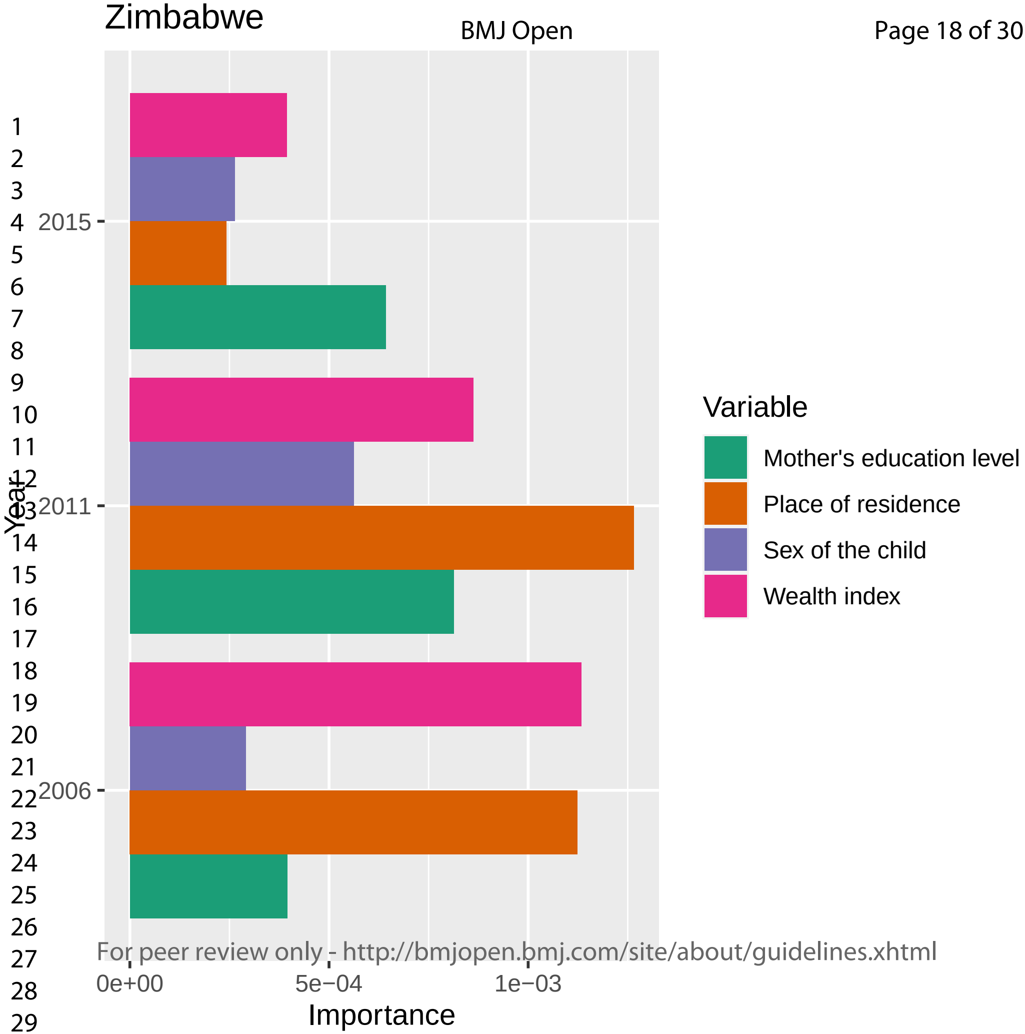
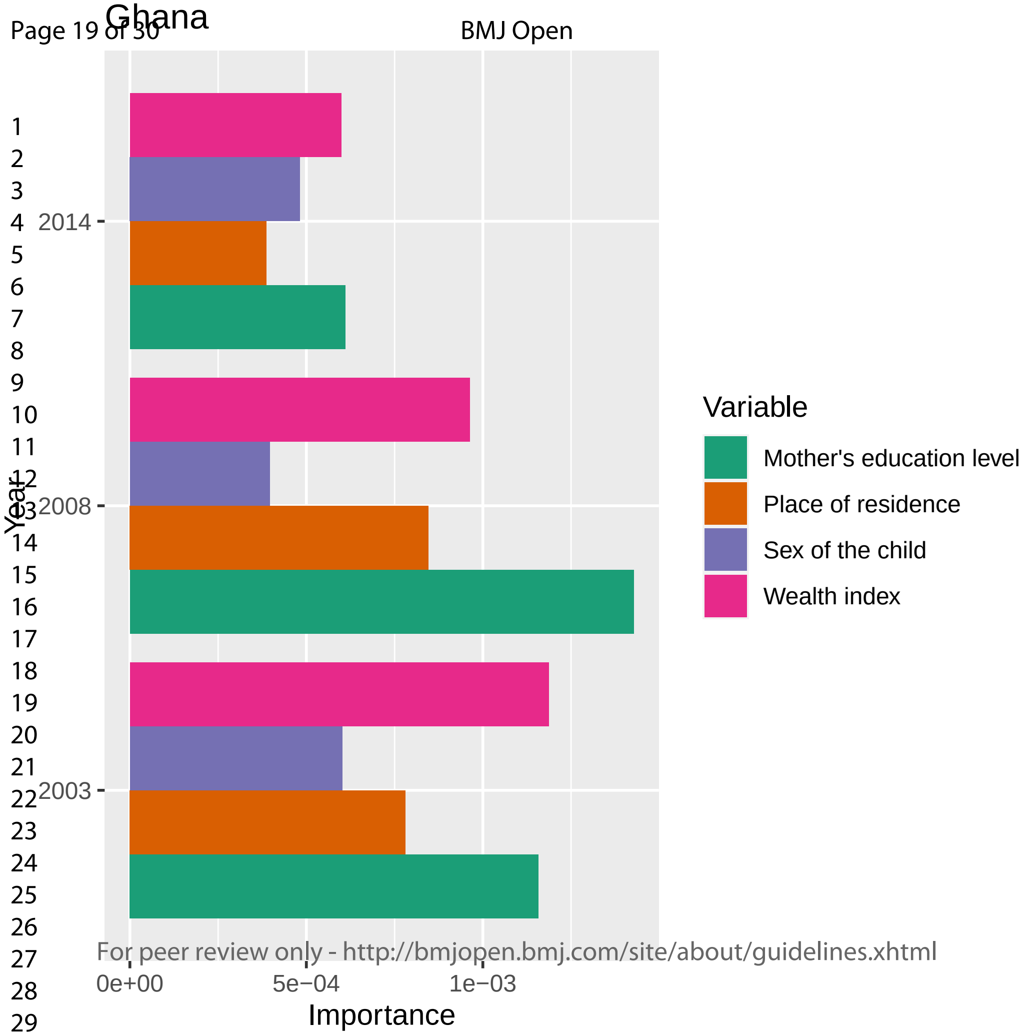
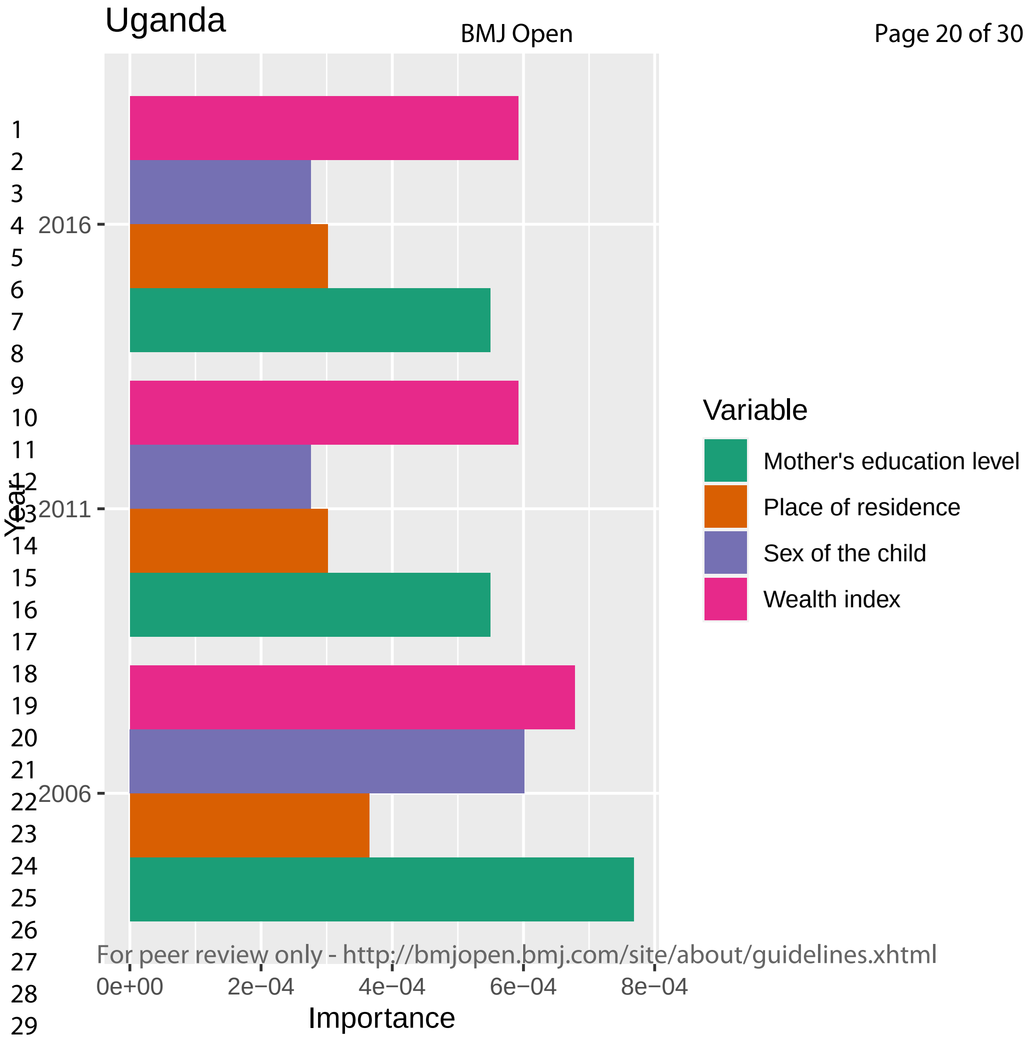


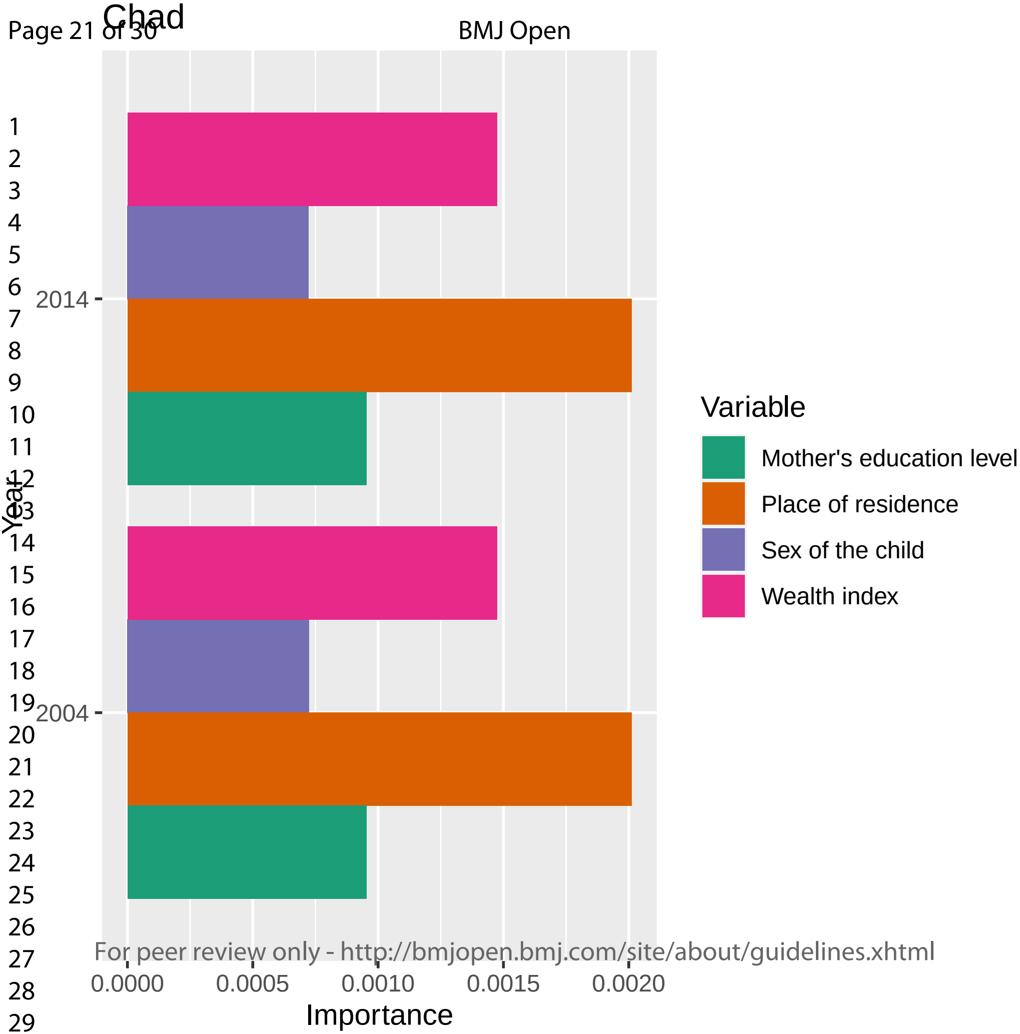
Fig 1. DeepSurv architecture Katzman et al.,[21]

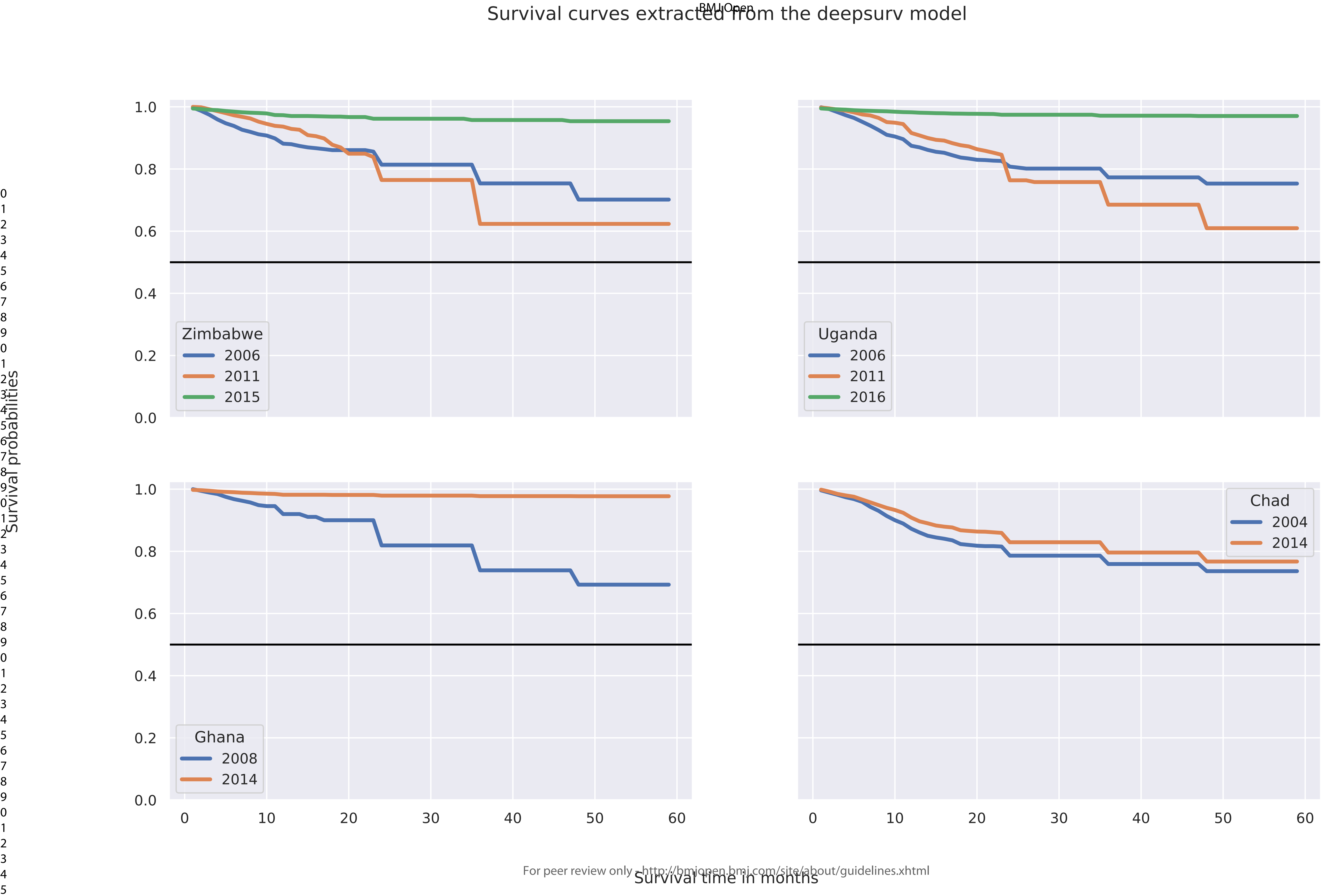
75x84mm (300 x 300 DPI)

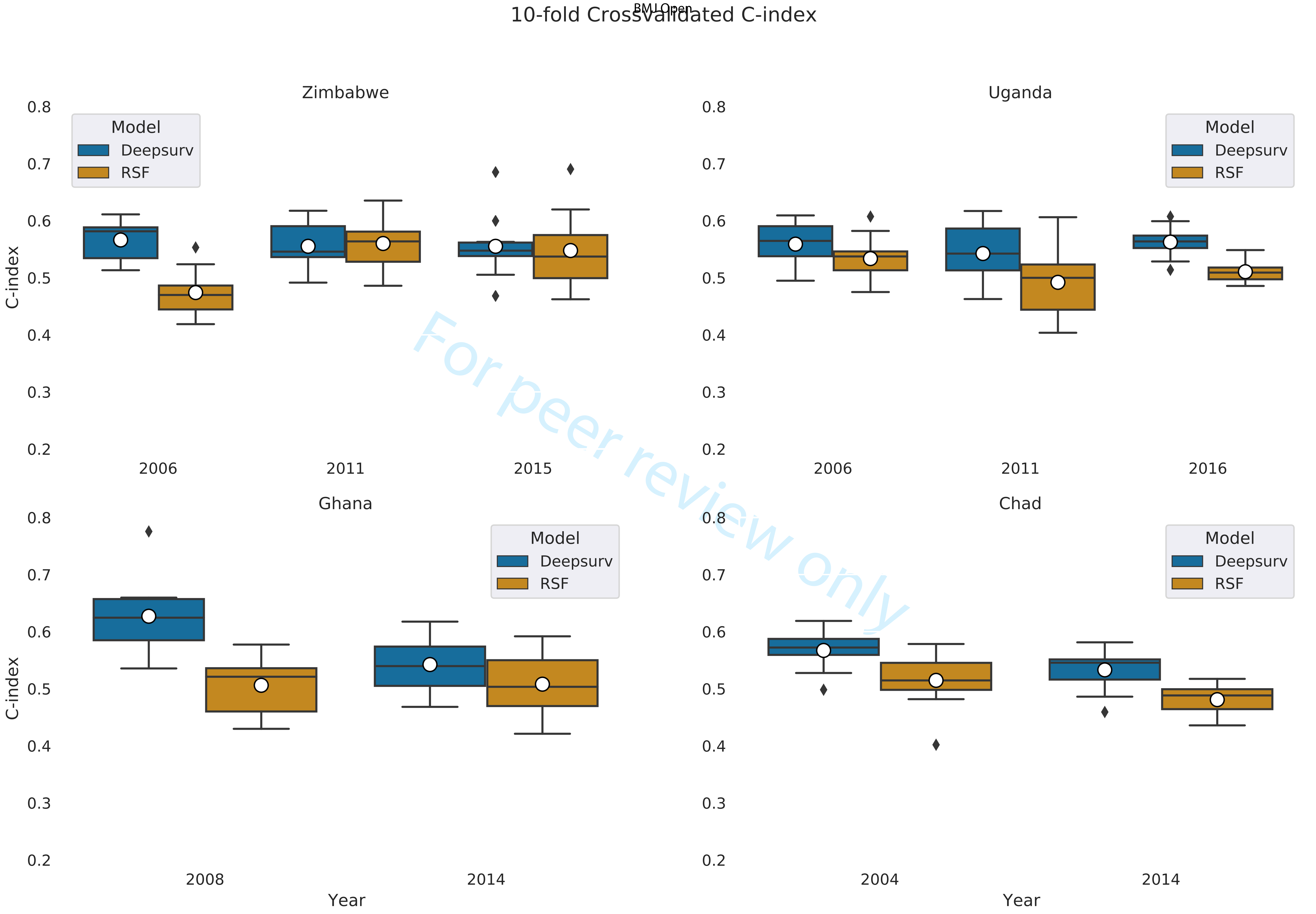

Zimbabwe
Year
2015
2011
2006
0e+00
5e−04
1e−03
Importance
Variable
Mother's education level
Place of residence
Sex of the child
Wealth index


Ghana
Year
2014
2008
2003
0e+00
5e−04
1e−03
Importance
Variable
Mother's education level
Place of residence
Sex of the child
Wealth index


Uganda
Year
2016
2011
2006
0e+00
2e−04
4e−04
6e−04
8e−04
Importance
Variable
Mother's education level
Place of residence
Sex of the child
Wealth index


Chad
Year
2014
2004
0.0000
0.0005
0.0010
0.0015
0.0020
Importance
Variable
Mother's education level
Place of residence
Sex of the child
Wealth index


Survival curves extracted from the deepsurv model
Survival probabilities
Survival time in months
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
40
50
60
Zimbabwe
2006
2011
2015
Uganda
2006
2011
2016
Ghana
2008
2014
Chad
2004
2014


10-fold Crossvalidated C-index
Zimbabwe
Uganda
Ghana
Chad
Model
Deepsurv
RSF
C-index
0.8
0.7
0.6
0.5
0.4
0.3
0.2
2006
2011
2015
2016
2008
2014
2004
Year

COREQ (COnsolidated criteria for REporting Qualitative research) Checklist

A checklist of items that should be included in reports of qualitative research. You must report the page number in your manuscript where you consider each of the items listed in this checklist. If you have not included this information, either revise your manuscript accordingly before submitting or note N/A.

Topic	Item No.	Guide Questions/Description	Reported on Page No.
Domain 1: Research team and reflexivity			
Personal characteristics			
Interviewer/facilitator	1	Which author/s conducted the interview or focus group?	DHS- 1 TO 4
Credentials	2	What were the researcher's credentials? E.g. PhD, MD	DHS- 1 TO 4
Occupation	3	What was their occupation at the time of the study?	N/A
Gender	4	Was the researcher male or female?	BOTH MALE AN
Experience and training	5	What experience or training did the researcher have?	DHS- 1 TO 4
Relationship with participants			
Relationship established	6	Was a relationship established prior to study commencement?	N/A
Participant knowledge of the interviewer	7	What did the participants know about the researcher? e.g. personal goals, reasons for doing the research	DHS- 1 TO 4
Interviewer characteristics	8	What characteristics were reported about the inter viewer/facilitator? e.g. Bias, assumptions, reasons and interests in the research topic	DHS- 1 TO 4
Domain 2: Study design			
Theoretical framework			
Methodological orientation and Theory	9	What methodological orientation was stated to underpin the study? e.g. grounded theory, discourse analysis, ethnography, phenomenology, content analysis	PAGE 1 TO 4
Participant selection			
Sampling	10	How were participants selected? e.g. purposive, convenience, consecutive, snowball	DHS- 1 TO 4
Method of approach	11	How were participants approached? e.g. face-to-face, telephone, mail, email	DHS- 1 TO 4
Sample size	12	How many participants were in the study?	TABLE 2 AND 3
Non-participation	13	How many people refused to participate or dropped out? Reasons?	DHS- 1 TO 4
Setting			
Setting of data collection	14	Where was the data collected? e.g. home, clinic, workplace	DHS- 1 TO 4
Presence of non-participants	15	Was anyone else present besides the participants and researchers?	DHS- 1 TO 4
Description of sample	16	What are the important characteristics of the sample? e.g. demographic data, date	PAGE 3 TO 5
Data collection			
Interview guide	17	Were questions, prompts, guides provided by the authors? Was it pilot tested?	DHS- 1 TO 4
Repeat interviews	18	Were repeat inter views carried out? If yes, how many?	N/A
Audio/visual recording	19	Did the research use audio or visual recording to collect the data?	NO
Field notes	20	Were field notes made during and/or after the inter view or focus group?	DHS- 1 TO 4
Duration	21	What was the duration of the inter views or focus group?	N/A
Data saturation	22	Was data saturation discussed?	N/A
Transcripts returned	23	Were transcripts returned to participants for comment and/or	N/A

Topic	Item No.	Guide Questions/Description	Reported on Page No.
		correction?	
Domain 3: analysis and findings			
Data analysis			
Number of data coders	24	How many data coders coded the data?	ONE
Description of the coding tree	25	Did authors provide a description of the coding tree?	NO
Derivation of themes	26	Were themes identified in advance or derived from the data?	DERIVED FROM
Software	27	What software, if applicable, was used to manage the data?	R AND PYTHON
Participant checking	28	Did participants provide feedback on the findings?	NO
Reporting			
Quotations presented	29	Were participant quotations presented to illustrate the themes/findings? Was each quotation identified? e.g. participant number	NO
Data and findings consistent	30	Was there consistency between the data presented and the findings?	YES
Clarity of major themes	31	Were major themes clearly presented in the findings?	YES
Clarity of minor themes	32	Is there a description of diverse cases or discussion of minor themes?	YES-PAGE 13

Developed from: Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. *International Journal for Quality in Health Care*. 2007. Volume 19, Number 6: pp. 349 – 357

Once you have completed this checklist, please save a copy and upload it as part of your submission. DO NOT include this checklist as part of the main manuscript document. It must be uploaded as a separate file.

Standards for Reporting Qualitative Research (SRQR)*

http://www.equator-network.org/reporting-guidelines/srqr/

Title and abstract

	Page/line no(s).
Title - Concise description of the nature and topic of the study Identifying the study as qualitative or indicating the approach (e.g., ethnography, grounded theory) or data collection methods (e.g., interview, focus group) is recommended	PAGE 1
Abstract - Summary of key elements of the study using the abstract format of the intended publication; typically includes background, purpose, methods, results, and conclusions	PAGE 1

Introduction

Problem formulation - Description and significance of the problem/phenomenon studied; review of relevant theory and empirical work; problem statement	PAGE 2
Purpose or research question - Purpose of the study and specific objectives or questions	PAGE 2

Methods

Qualitative approach and research paradigm - Qualitative approach (e.g., ethnography, grounded theory, case study, phenomenology, narrative research) and guiding theory if appropriate; identifying the research paradigm (e.g., postpositivist, constructivist/ interpretivist) is also recommended; rationale**	PAGE 5 TO 9
Researcher characteristics and reflexivity - Researchers' characteristics that may influence the research, including personal attributes, qualifications/experience, relationship with participants, assumptions, and/or presuppositions; potential or actual interaction between researchers' characteristics and the research questions, approach, methods, results, and/or transferability	PAGE 9 TO 11
Context - Setting/site and salient contextual factors; rationale**	
Sampling strategy - How and why research participants, documents, or events were selected; criteria for deciding when no further sampling was necessary (e.g., sampling saturation); rationale**	PAGE 1 TO 5
Ethical issues pertaining to human subjects - Documentation of approval by an appropriate ethics review board and participant consent, or explanation for lack thereof; other confidentiality and data security issues	PAGE 13
Data collection methods - Types of data collected; details of data collection procedures including (as appropriate) start and stop dates of data collection and analysis, iterative process, triangulation of sources/methods, and modification of procedures in response to evolving study findings; rationale**	PAGE 1 TO 5 AND PAGE 13

Data collection instruments and technologies - Description of instruments (e.g., interview guides, questionnaires) and devices (e.g., audio recorders) used for data collection; if/how the instrument(s) changed over the course of the study	N/A
Units of study - Number and relevant characteristics of participants, documents, or events included in the study; level of participation (could be reported in results)	PAGE 13
Data processing - Methods for processing data prior to and during analysis, including transcription, data entry, data management and security, verification of data integrity, data coding, and anonymization/de-identification of excerpts	N/A
Data analysis - Process by which inferences, themes, etc., were identified and developed, including the researchers involved in data analysis; usually references a specific paradigm or approach; rationale**	PAGE 9 TO 11
Techniques to enhance trustworthiness - Techniques to enhance trustworthiness and credibility of data analysis (e.g., member checking, audit trail, triangulation); rationale**	PAGE 13

Results/findings

Synthesis and interpretation - Main findings (e.g., interpretations, inferences, and themes); might include development of a theory or model, or integration with prior research or theory	PAGE 11 TO 12
Links to empirical data - Evidence (e.g., quotes, field notes, text excerpts, photographs) to substantiate analytic findings	PAGE 11 TO 12

Discussion

Integration with prior work, implications, transferability, and contribution(s) to the field - Short summary of main findings; explanation of how findings and conclusions connect to, support, elaborate on, or challenge conclusions of earlier scholarship; discussion of scope of application/generalizability; identification of unique contribution(s) to scholarship in a discipline or field	PAGE 11 TO 12
Limitations - Trustworthiness and limitations of findings	PAGE 1

Other

Conflicts of interest - Potential sources of influence or perceived influence on study conduct and conclusions; how these were managed	PAGE 13
Funding - Sources of funding and other support; role of funders in data collection, interpretation, and reporting	PAGE 13

*The authors created the SRQR by searching the literature to identify guidelines, reporting standards, and critical appraisal criteria for qualitative research; reviewing the reference lists of retrieved sources; and contacting experts to gain feedback. The SRQR aims to improve the transparency of all aspects of qualitative research by providing clear standards for reporting qualitative research.

**The rationale should briefly discuss the justification for choosing that theory, approach, method, or technique rather than other options available, the assumptions and limitations implicit in those choices, and how those choices influence study conclusions and transferability. As appropriate, the rationale for several items might be discussed together.

Reference:

O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. **Standards for reporting qualitative research: a synthesis of recommendations.** *Academic Medicine*, Vol. 89, No. 9 / Sept 2014
DOI: 10.1097/ACM.0000000000000388

STROBE Statement—checklist of items that should be included in reports of observational studies

	Item No.	Recommendation	Page No.	Relevant text from manuscript
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract	1	Abstract
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found	1	Abstract
Introduction				
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported	1-2	Introduction
Objectives	3	State specific objectives, including any prespecified hypotheses	2	Introduction
Methods				
Study design	4	Present key elements of study design early in the paper	3	Data
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	3-5	Data
Participants	6	*(a)* *Cross-sectional study*—Give the eligibility criteria, and the sources and methods of selection of participants	3-5	Data
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable	4	Data and Data pre- processing
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	*1-5*	Introduction and Data
Bias	9	Describe any efforts to address potential sources of bias	3- 5	Data and Data pre- processing
Study size	10	Explain how the study size was arrived at	1	Abstract

Continued on next page

Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why	4 -5	Data pre-processing
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding	5-9	Models
		(*b*) Describe any methods used to examine subgroups and interactions		N/A
		(*c*) Explain how missing data were addressed	4-5	Data pre-processing
		(*d*) *Cross-sectional study*—If applicable, describe analytical methods taking account of sampling strategy		N/A
		(*e*) Describe any sensitivity analyses		N/A
Results				
Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed	3-5	Data
		(b) Give reasons for non-participation at each stage	3-5	Data
		(c) Consider use of a flow diagram		N/A
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders	Table 2 and 3	Page 3 and 5
		(b) Indicate number of participants with missing data for each variable of interest	N/A	
		(c) *Cohort study*—Summarise follow-up time (eg, average and total amount)		
Outcome data	15*	*Cohort study*—Report numbers of outcome events or summary measures over time		
		Case-control study—Report numbers in each exposure category, or summary measures of exposure		
		Cross-sectional study—Report numbers of outcome events or summary measures	3- 5	Data and Data pre- processing
Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included	9-11	Results
		(*b*) Report category boundaries when continuous variables were categorized	N/A	
		(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	N/A	

Continued on next page

Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses	N/A	
Discussion				
Key results	18	Summarise key results with reference to study objectives	11-12	Discussion
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	11-12	Discussion
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	11-12	Discussion
Generalisability	21	Discuss the generalisability (external validity) of the study results	12	Conclusion
Other information				
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	13	Acknowledgement

*Give information separately for cases and controls in case-control studies and, if applicable, for exposed and unexposed groups in cohort and cross-sectional studies.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article