A PROGNOSTIC MODEL TO PREDICT SURVIVAL IN CHILDREN WITH EBOLA VIRUS DISEASE

A Master’s Thesis Presented

By

KELSEY BUTLER

Submitted to the Faculty of the Morningside Graduate School of Biomedical Sciences at UMass Chan Medical School
In partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

NOVEMBER 8, 2022

BIOINFORMATICS AND COMPUTATIONAL BIOLOGY
A PROGNOSTIC MODEL TO PREDICT SURVIVAL IN CHILDREN WITH EBOLA VIRUS DISEASE

A Master’s Thesis Presented

By

KELSEY BUTLER

The signatures of the Master’s Thesis Committee signify Completion and approval as to style and content of the Thesis

Dr. Vanni Bucci, Chair of Committee

Dr. Elinor Karlsson, Member of Committee

Dr. Adrian Zai, Member of Committee

The signature of the Dean of the Morningside Graduate School of Biomedical Sciences signifies that the student has met all Master’s Degree graduation requirements of the school.

Mary Ellen Lane, Ph.D.,
Dean of the Graduate School of Biomedical Sciences

Bioinformatics and Computation Biology

November 8, 2022
ACKNOWLEDGMENTS

I would like to express my deepest appreciation to my advisor, Dr. Andrés Colubri, for his guidance on this analysis. Completion of this project would not have been possible without Drs. Ian Michelow and Alicia Genisca, as well as Monique Gainey, who spent many hours in Zoom meetings discussing models and data. Thanks should also go to all authors on the final published manuscript for editing and feedback throughout the peer review process. Lastly, I would be remiss in not mentioning the clinicians and providers who worked on-the-ground to treat patients and collect data, risking their own safety for the advancement of medicine.
ABSTRACT

Repeated outbreaks of Ebola Virus Disease (EVD) in low-resource settings emphasize the importance of evidence-based guidelines to direct treatment. Previous research has shown that EVD causes high case fatality rates (CFRs) in young children, yet there are limited data focusing on pediatric patients. Here we present a prognostic model to predict mortality in children who are Ebola-positive using information available during the first 48 hours after admission to the treatment center. A logistic regression model was trained on triage data from the Ebola Data Platform, a repository of retrospective patient data compiled from actors that responded to the West African EVD outbreak from 2014-2016. Patients <18 years of age were included in the analysis (N=579) and the CFR was 40%. Overall 13% of data were missing, and multiple imputation was used to estimate missing values. Variable selection using elastic net regularization selected age, CT value, bleeding, breathlessness, bone or muscle pain, anorexia, swallowing problems, and diarrhea as predictors. Bootstrap validation yielded an optimism-corrected area under the curve (AUC) of 0.75 (95% CI: 0.71-0.79). The model was externally validated using data from the current EVD outbreak in the Democratic Republic of the Congo (DRC). While the model's discriminative ability on the DRC data was similar (AUC=0.75, 95% CI: 0.63-0.87) to the training data, calibration was poor. We recalibrated the model by re-estimating the intercept and slope, and further improved model performance by including aspartate aminotransferase (AST) as a biomarker. The
updated model with AST as an added predictor has an AUC of 0.90 (95% CI: 0.77-1). These preliminary results are encouraging but should be interpreted with caution because of limited availability of AST values in the validation data (n=25). The prognostic model described here has promising potential for use in a clinical setting and will continue to be validated as more data becomes available. Future efforts will focus on integrating the validated model into mHealth tools to aid clinicians in making informed, data-driven decisions about patient care.
# TABLE OF CONTENTS

List of tables ............................................................................................................... vii
List of figures ................................................................................................................. viii
List of third party copyrighted material ................................................................. ix
List of abbreviations .................................................................................................. x
Preface ......................................................................................................................... xi

CHAPTER 1: INTRODUCTION ................................................................. 1
  1.1 2014-2016 Ebola outbreak in West Africa ................................................ 1
  1.2 2018–2020 Kivu Ebola epidemic in DRC ................................................... 3
  1.3 Virology and clinical presentation of EVD ................................................. 5
  1.4 Differential risk and disease presentation for children under age five .... 7

CHAPTER 2: METHODOLOGY .......................................................... 10
  2.1 Preliminary analysis ................................................................................. 10
  2.2 Overview of logistic regression ............................................................... 11
  2.3 Descriptive data analysis ......................................................................... 12
  2.4 Multiple imputation ............................................................................... 13
  2.5 Variable selection .................................................................................. 14
  2.6 Model fitting ........................................................................................... 16
  2.7 Model evaluation ................................................................................... 17
  2.8 Model update ......................................................................................... 20

CHAPTER 3: RESULTS ....................................................................... 22
  3.1 Summary of West Africa data ................................................................. 22
  3.2 Summary of DRC data ........................................................................... 27
  3.3 Model selection and performance ............................................................. 29
  3.4 Model update ......................................................................................... 35

CHAPTER 4: DISCUSSION ............................................................... 39
  4.1 Overview of contribution to the field ....................................................... 39
  4.2 Limitations of research ........................................................................... 40
  4.3 Conclusion ............................................................................................... 42

REFERENCES .................................................................................... 43

APPENDIX A: DATA COLLECTION METHODOLOGY ......................... 47
**LIST OF TABLES**

Table 3.1 Demographic and clinical characteristics of patients in the West Africa derivation cohort.......................................................... 22

Table 3.2 Summary of missing values in West Africa derivation dataset........ 25

Table 3.3 Demographic and clinical characteristics of patients in DRC validation cohort ........................................................................................................... 27

Table 3.4 Comparison of baseline characteristics in West Africa derivation and DRC validation cohorts .................................................................................................................. 29

Table 3.5 Results of variable selection protocol for categorical variables........ 30

Tables 3.6a & 3.6b Confusion matrix and detailed performance measures of the EPiC model................................................................. 32

Tables 3.7a & 3.7b Confusion matrix and detailed performance measures of the EPiC model augmented with AST and CK........................................... 37
LIST OF FIGURES

Figure 3.1 Map of ETU locations .......................................................... 24
Figure 3.2 Distribution of imputed CT values ............................................. 26
Figure 3.3 Performance characteristics of the prediction model ............... 31
Figure 3.4 Receiver operating characteristic (ROC) and calibration plots for the minimal model ................................................................. 34
Figure 3.5 Plots showing the importance of the features in the EPIC model .... 35
Figure 3.6 Discrimination and calibration curves for model with additional biomarker ................................................................. 38
LIST OF THIRD PARTY COPYRIGHTED MATERIAL

Figure 3.1 was plotted with the R package tmap,\(^1\) using base layer maps in the public domain from the Natural Earth project

(https://www.naturalearthdata.com/about/terms-of-use/).
## LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALT</td>
<td>Alanine transaminase</td>
</tr>
<tr>
<td>AST</td>
<td>Aspartate transaminase</td>
</tr>
<tr>
<td>AUC</td>
<td>Area under the ROC curve</td>
</tr>
<tr>
<td>CFR</td>
<td>Case-fatality rate</td>
</tr>
<tr>
<td>CK</td>
<td>Creatine kinase</td>
</tr>
<tr>
<td>DRC</td>
<td>Democratic Republic of the Congo</td>
</tr>
<tr>
<td>EPiC</td>
<td>Ebola Virus Disease Prognosis in Children model</td>
</tr>
<tr>
<td>ETU</td>
<td>Ebola treatment unit</td>
</tr>
<tr>
<td>EVD</td>
<td>Ebola virus disease</td>
</tr>
<tr>
<td>IMC</td>
<td>International Medical Corps</td>
</tr>
<tr>
<td>IQR</td>
<td>Interquartile range</td>
</tr>
<tr>
<td>ROC</td>
<td>Receiver operating characteristic</td>
</tr>
</tbody>
</table>
PREFACE

The models described herein are published and accessible at https://doi.org/10.1371/journal.pntd.0010789. All writing in this thesis was completed by the author, Kelsey M. Butler, with the exception of Appendix A: Data Collection Methodology, which was provided by Monique Gainey.
Chapter 1

INTRODUCTION

Chapter 1.1 2014-2016 Ebola outbreak in West Africa

Geography and spread

The 2014-2016 Ebola Virus Disease (EVD) outbreak in West Africa was the largest and most deadly outbreak since the disease was first recorded in 1976.\textsuperscript{1-3} By the time it was declared over in June 2016, the total case count and death toll surpassed that of all previous outbreaks.\textsuperscript{1,4} The widespread impact can be attributed to a unique confluence of factors including a large geographical distribution of cases, increased migration and travel, disease spread in cities, and a delayed public health response.\textsuperscript{1-3}

Ebola Virus Disease is caused by the Ebola virus, which takes natural reservoir in bats, and is spread by zoonotic transmission when a person contacts blood or body fluids of an infected animal.\textsuperscript{1} A retrospective analysis traced the index case for the 2014 outbreak to an 18-month-old male child in Guinea who was likely infected by bats in December 2013.\textsuperscript{2} The unusual symptom presentation allowed sustained community spread and the first case was not publicly reported until March 2014 in Guinea.\textsuperscript{1} Genetic analysis supports this narrative, and showed that the West Africa variant initiated in Guinea then spread to Sierra Leone in May 2014.\textsuperscript{5} By early summer 2014, the outbreak had spread to major cities in three countries (Guinea, Liberia, Sierra Leone), which contributed to rapid transmission and quickly overwhelmed areas with limited
resources, prompting an international response. Previous outbreaks had not reached cities, and public health authorities were unprepared for the magnitude of response required to combat largescale transmission of EVD. By mid-2014, the outbreak was growing exponentially with cases doubling every 34.8 days. In addition to the three countries with the highest caseload, cases were introduced in seven other countries which didn’t ultimately see substantial spread.

**Public response**

The major actors participating in the public response were local and international governments, the World Health Organization, and non-governmental organizations. In Sierra Leone, community engagement and social mobilization was key to the response and included data collection, education campaigns on preventing transmission (safe burials, getting tested when symptomatic), and actions to limit transmission, i.e. installing hand washing stations. This local response was led by community activists, religious leaders, and radio broadcasters. Implementing strategies to mitigate transmission, treat infected persons, and collect data all require community cooperation and engagement, and in some settings clinical trials involving convalescent plasma and vaccines were inhibited by lack of trust in health workers. Ebola Treatment Units (ETUs) run by public agencies can provide a safe place to test, treat, and isolate infected people. In September 2014, International Medical Corps (IMC)
opened five ETUs in Sierra Leone and Liberia that focused on treating suspected cases and collecting clinical and laboratory data.\(^8\)

**Impact**

The final toll of the outbreak exceeded 11,000 deaths and 28,600 confirmed cases.\(^3,6,7\) During the outbreak, 8% of healthcare workers in Liberia and 7% in Sierra Leone were killed by Ebola.\(^9\) This contributed to government healthcare systems in Guinea, Liberia, and Sierra Leone becoming quickly overwhelmed as the disease spread and access to other healthcare services was limited, increasing deaths from other common illnesses.\(^10\) The economic cost of the outbreak was 4.3 billion US dollars, and the hardest hit countries saw reduced economic activity, food insecurity, and reduced travel and trade.\(^1\) The long-term mental health effects on survivors, healthcare workers, and response staff was substantial, with one study reporting that 20% of respondents met criteria for Post-Traumatic Stress Disorder.\(^11\) This and other retrospective analyses of the outbreak were possible because the timing of the outbreak coincided with largescale data collection efforts, including mobile health applications,\(^12\) which facilitated recording large and diverse datasets.

*Chapter 1.2 2018–2020 Kivu Ebola epidemic in the Democratic Republic of the Congo (DRC)*

*Geography and spread*
The 2018–2020 Kivu Ebola epidemic was the second largest Ebola outbreak ever recorded. The first cluster of cases was observed in North Kivu in July 2018, and additional cases were reported in Ituri the next month. A complicating factor in this outbreak was its occurrence in an active conflict zone, in which armed groups were present in affected areas during a time of contentious elections. Similar to the West Africa outbreak, there was significant migration and travel to and from affected area and high risk of wider geographical spread. Unlike the earlier outbreak, transmission was largely contained to North Kivu and Ituri. This can be attributed to screening for infection, contact tracing, lab testing of samples, vaccination, and treatment. This rapid and effective response is in marked contrast to the 2014-2016 epidemic.

Public response and impact

The World Health Organization coordinated with DRC Ministry of Health to train healthcare workers, perform contact tracing, confirm lab diagnoses, and treat and vaccinate patients. As in previous outbreaks, local community engagement was key in preventing the spread. Highly effective vaccines were administered to over 300,000 people, which required large-scale coordination and trust of HCW. The response to this outbreak integrated information learned from previous responses. The public health response included field laboratories to test samples, Ebola Treatment Units (ETUs) to care for patients, and centers to isolate suspected cases. It is notable expanded treatment options were
available during the outbreak. An Ebola vaccine was licensed and two treatments were available—Regeneron (REGN-EB3) and mAb114. The outbreak was declared over in June 2020, 42 days after the last known infected person tested negative. In total, there were 3481 cases recorded and 2299 deaths, giving a case-fatality rate of 66%. Of all the cases recorded, 29% occurred in children.

Chapter 1.3 Virology and clinical presentation of EVD

Virology

Ebola virus is in the Filoviridae family of negative strand RNA viruses which cause viral hemorrhagic fevers in humans and other animals. Of the five Ebola virus subtypes, Zaire ebolavirus is the most lethal, with a 50-90% mortality rate, and was the infecting species of the 2014-2016 epidemic. In the viral genome, seven open reading frames produce eight proteins. The viral lifecycle begins with viral entry mediated by transmembrane glycoprotein GP. Furin-mediated cleavage of GP produces GP1 and GP2, linked by a disulfide bond. GP1 initiates attachment to the host cell followed by GP2 led fusion of viral and host membranes. Once the virus has entered the cell, uncoating and fusion of viral and host endosomal membranes occurs. Next the viral polymerase transcribes mRNA, followed by translation, which requires formation of a ribonucleoprotein (RNP). Finally, assembly of virions and budding at plasma membrane complete the life cycle. Transmembrane GP facilitates entry into monocytes, macrophages, and endothelial cells. Collectively, this contributes
to hemorrhagic fever by targeting cells that line blood vessels and inducing cell damage which leads to release of inflammatory of fever-inducing cytokines.\textsuperscript{19}

\textit{Symptoms and transmission}

Patients infected with EVD initially present with nonspecific symptoms, including fever, fatigue, body aches, and headache. As the disease progresses, patients experience vomiting, diarrhea, and bleeding, and in the most severe cases this progresses to multiorgan failure and death.\textsuperscript{1,19} Human-to-human transmission occurs with contact of blood or body fluids, and once a person has been exposed, the incubation period is 2-21 days. The virus multiplies rapidly and is thus able to overwhelm the host humoral and cellular immune response, leading to high viral loads.\textsuperscript{19} Infectivity corresponds to viral load and increases daily with disease progression.\textsuperscript{8} Deceased persons who reached end-stage disease have the highest viral loads, and cultural burial practices that involve washing and physical contact with corpses contribute to spread.\textsuperscript{1}

\textit{Treatment}

Outbreaks of EVD have case-fatality rates (CFRs) of 25-90% and occur in low resource settings.\textsuperscript{1} Death is due to dehydration from severe loss of body fluids (diarrhea, vomiting, hemorrhage) leading to low blood pressure and hypovolemic shock.\textsuperscript{4} Supportive care is one of the most important modes of treatment, consisting of administration of fluids and electrolytes orally and
intravenously to combat dehydration, and analgesics for pain relief. Early and aggressive hydration accounts for a four-fold difference in mortality in lower to middle income countries.

Vaccines for EVD are available, though guidelines and protocols for vaccination are still in development and generally suggest that vaccines be administered in the event of an outbreak. Pharmacological treatments include Remdesivir which was developed to treat EVD and implemented as an experimental treatment during the 2014-2016 outbreak. Trial results showed it to be ineffective in treating Ebola, and it was later repurposed to treat COVID-19. Inmazeb by Regeneron combines three human monoclonal antibodies that all target the Ebola virus GP. This treatment was not available during the 2014-2016 outbreak but was using during the 2018-2020 DRC outbreak. Additional monoclonal antibody treatments are Mab 114, a single monoclonal antibody that targets GP, and ZMAPP, a combo of 3 monoclonal antibodies.

Chapter 1.4 Differential risk and disease presentation for children under age five

The index cases for both the West Africa and DRC outbreaks are believed to have been children. Children are affected less often than adults and usually comprise only about 10% of cases, possibly due to lower rates of exposure because they are not caring for sick relatives. In recent outbreaks, an increasing proportion of EVD cases in children have been observed—in the West
Africa outbreak, 18-20% of all cases were children, and that number rose to 29% in the DRC outbreak. Research into EVD has mostly aggregated data from all ages, despite the observation that EVD manifests in children differently than in adults. Young children are especially vulnerable and have the shortest incubation period of any age group (6.9 days for children less than 1 year of age, 10 days for children 10-15 years of age), more rapid disease progression, and a high CFR, with the highest rates in children under four years of age.

Symptoms at presentation are also distinct, with children less likely to present with fever than adults, which highlights the difficulty of early diagnosis in this age group.

Research focusing on EVD in children is limited and studies suffer from small sample sizes. One research group analyzed data from 91 children under five years old who were infected in the Sierra Leone 2014 EVD outbreak. They reported that children under two years old had the highest CFR, and the most common presenting symptoms were weakness, fever, distress, anorexia, diarrhea, and cough. In this study, 25% of children did not present with fever. Another study using data from the same outbreak expanded the inclusion criteria to include children under thirteen years old (N = 309). Here they reported younger age and diarrhea were risk factors for mortality, and when death occurred it was rapid, occurring on average just three days after admission. This conclusion was supported by an additional study reporting that most deaths in children occur within ten days of admission (N = 33). In the 2000-2001
Northern Uganda outbreak, only 9% of cases were kids and CFR for children was 40, compared to 50 for adults. In this study, all pediatric cases presented with fever. As in other reports, children under five years old had highest CFR, in this case 80%. The high CFR for young children was first observed in the first EVD outbreak in 1976, when 11 neonates born to mothers with EVD all died.
CHAPTER 2

METHODOLOGY

Chapter 2.1 Preliminary Analysis

Prior to data exploration and model building, patterns of missing data were analyzed, and the dataset was prepared for subsequent analyses. The use of machine learning algorithms requires a complete set of data with no missing values. This was achieved by using multiple imputation to fill in missing values, a protocol which itself requires that data be missing completely at random (MCAR). Data are MCAR if the missing values are unrelated to individual attributes of each case. To test if data were MCAR, Little’s MCAR test was performed using the naniar package in R. Little’s test uses the $\chi^2$ statistic to test the null hypothesis that there are no differences between means of different missing patterns. In this instance, Little’s test was applied separately to data from each ETU to test patterns of missingness by collection site.

The Shapiro-Wilk test for normality was performed to test if the data was drawn from a normally distributed population (Equation 3.1). In this equation, the numerator corresponds to the slope of a plot of quantiles of observed data versus expected quantiles of data drawn from a normal distribution with the same mean and standard deviation as the sample. The denominator represents the population variance. If the data are normally distributed, the test statistic $W$ equals 1.
Shapiro-Wilk test:

\[
W = \frac{(\sum_{i=1}^{n} a_i x_i)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}
\]

To determine the upper limit on number of predictors in the model, we applied \( p < \frac{m}{15} \), where \( p \) is number of predictors and \( m \) is the limiting sample size.\(^3\) In the scenario with a binary outcome variable of death or survival, the value of \( m \) is equivalent to the minimum number of total deaths and total surviving cases. Finally, variables with greater than 50% of values missing were excluded from analysis.

**Chapter 2.2 Overview of Logistic Regression**

The logistic regression model is a logistic transformation of the linear predictor \( X\beta = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k \).\(^3\) It models the probability that a binary outcome variable \( Y = 1 \) given the observed data \( X \):

\[
p = \text{Prob}(Y = 1|X) = \frac{1}{1 + e^{-x\beta}}
\]

\[
\text{logit}(p) = \log \left( \frac{p}{1-p} \right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k
\]

This function restricts the value of the response variable to the range 0 to 1. The key assumptions of logistic regression include: (1) observations are independent, (2) no multicollinearity between variables, and (3) independent variables are linearly related to log odds.
Parameter estimation

Maximum likelihood estimation was used to estimate parameters. The likelihood function $L(\theta)$ describes the likelihood of a set of parameter values $\theta$ given the observed data. The general form is:

$$L(\theta) = p(y \mid X; \theta)$$

In binary logistic regression, the joint likelihood for $N$ observations is the product of likelihoods for each individual case. In this formula, $P_i$ is the predicted probability and $Y_i$ is the observed outcome.

$$L(\theta) = \prod_{i=1}^{N} P_i^{Y_i}[1 - P_i]^{1-Y_i}$$

To simplify the calculations, we use the log likelihood function as the final form which we seek to maximize using an iterative algorithm implemented with the `lrm` function from the `rms` package.

$$L(\theta) = \sum_{i=1}^{N} Y_i \log P_i + (1 - Y_i) \log (1 - P_i)$$

Chapter 2.3 Descriptive Data Analysis

If the Shapiro-Wilk test for normality indicated that data were not normally distributed, variables were analyzed by calculating median and interquartile range (IQR) values. Binary symptom variables were analyzed by calculating total incidence in patients who survived or died.
Odds ratios

The odds ratio describes the odds of survival given the presence of a specific disease characteristic. Odds ratios and p-values for binary variables were calculated from univariate regression coefficients. For continuous outcome variables, odds ratios are reported for a five-year increase in age and for an increase in Ct by IQR.

Logistic regression model for a single binary predictor $X$

$$\text{logit}(Y = 1|X = 0) = \beta_0$$

$$\text{logit}(Y = 1|X = 1) = \beta_0 + \beta_1$$

$$\text{Odds ratio} = e^{\beta_1}$$

Chapter 2.4 Multiple Imputation

In the West Africa data, 10.7% of values were missing for 16 of 18 predictor variables, and multiple imputation was used to address missing data. We used the aregImpute function from the R package Hmisc to create 100 imputed datasets. In this approach, each imputation is generated by fitting a flexible additive model on a unique bootstrap resample drawn from the original data. We fit a saturated imputation model using all 18 candidate predictors and allowed for nonlinear modeling of continuous variables using restricted cubic splines.
Chapter 2.5 Variable Selection

For variable selection, we opted to use Elastic Net regularization, which combines the Lasso and Ridge regression methods. This technique is effective in handling multicollinearity and selects variables for a parsimonious model that is more generalizable to the population.\textsuperscript{41,42} Eighteen candidate predictors including age, sex, Ct value, and fifteen other epidemiological and clinical variables based on the current WHO criteria for identifying suspected Ebola cases were selected for inclusion in the model.\textsuperscript{43,44} These variables included fever, headache, respiratory distress (defined as fast respiratory rate; nasal flaring, grunting, intercostal recession and tracheal tug; in-drawing of lower chest wall; central cyanosis of lips and tongue; inability to breastfeed or drink; lethargy), bone or muscle pain, joint pain, conjunctivitis, asthenia, abdominal pain, hiccups, unexplained bleeding, vomiting, diarrhea, nausea, anorexia, or dysphagia.\textsuperscript{43,44}

\textit{Elastic net regularization}

Regularized regression is a technique to constrain or limit the magnitude of the coefficient estimates, which reduces model complexity and the chance of overfitting. Regularization adds a penalty term to the Residual Sum of Squares (RSS) cost function of linear regression. This equation describes the RSS for an $M \times p$ matrix with parameters $w$:

$$RSS = \sum_{i=1}^{M} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{M} \left( y_i - \sum_{j=0}^{p} w_j \times x_{ij} \right)^2$$
Regularized regression adds a penalty term to the RSS cost function. Two common regularization techniques are Lasso and Ridge regression. Lasso regression uses the L1 norm to penalize the absolute values of the coefficients, a technique that removes irrelevant variables by setting their coefficients to zero. Ridge regression uses the L2 norm to penalize the square of the coefficients, thus shrinking the size of coefficients toward zero while retaining them in the model. In each case, the hyperparameter $\lambda$ controls the strength of the penalty.

\[ Lasso \quad \text{(L1)} \quad \lambda \sum_{j=0}^{p} |w_j| \]
\[ Ridge \quad \text{(L2)} \quad \lambda \sum_{j=0}^{p} w_j^2 \]
\[ Elastic \quad net \quad penalty \quad \lambda \left( \frac{1 - \alpha}{2} \sum_{j=0}^{p} w_j^2 + \alpha \sum_{j=0}^{p} |w_j| \right) \]

Elastic net linearly combines the Lasso and Ridge regression penalties to overcome the limitations of each. In Elastic net, an additional hyperparameter $\alpha$ is introduced, in which $\alpha = 0$ specifies Ridge regression, $\alpha = 1$ specifies Lasso regression, and $0 < \alpha < 1$ combines them to reduce the magnitude of some coefficients and eliminate others by setting them to zero.
Variable selection protocol using Elastic net

The variable selection protocol worked as follows: Elastic Net was applied to each imputed dataset, the sign of the coefficients of the binary symptom variables in the resulting models were tallied, and those variables with the percentage of positive model coefficients above a given threshold were selected. This selection criterion facilitated the inclusion of groups of correlated predictors and predictors with small but significant effects. \(^{41}\) The threshold for variable inclusion was set at 100\% to exclude variables with weak and/or inconsistent effects.

The intermediate models generated during variable selection procedure were fitted with the glmnet() function in the R package glmnet.\(^{42}\) Continuous variables were modeled as nonlinear terms using restricted cubic splines with three knots. The logistic regression model was fit separately on 100 imputed datasets and the coefficients were averaged to produce a final model using the lrm and fit.mult.impute functions in the R packages rms\(^{45}\) and Hmisc\(^{40}\), respectively.

Chapter 2.6 Model Fitting

A saturated model was constructed to serve as a baseline against which to compare the performance of other predictive models. The model included age and Ct value as continuous predictors along with four binary symptom variables selected with the Elastic Net as described above. Bootstrap resampling was used
for internal validation. Bootstrapping is a resampling technique to generate several simulated datasets from a single sample. In this scheme, the sampling distribution is estimated by resampling \( n \) observations from the observed data of size \( n \) with replacement. This creates multiple samples of size \( n \), drawn from the original data, that approximate additional samples drawn from the population. This provides a way to estimate standard errors and confidence intervals without having to collect additional samples from the population.

Chapter 2.7 Model Evaluation

Discrimination

Model discrimination is a measure of how well the model distinguishes between high and low risk cases. This is used as the initial assessment of model fit.\(^{46}\) In binary logistic regression, discrimination is evaluated from the receiver operating characteristic (ROC) curve. The ROC curve plots sensitivity (true positive rate) versus 1 - specificity (false positive rate). A perfect model is one that correctly predicts the outcome for all patients with zero misclassifications, which produces an area under the curve (AUC) equal to one. A model which fails to discriminate between high and low risk patients has an AUC = 0.5, indicating predictions that are no better than random guessing. Higher AUC values correspond to better predictive models and are a first-line tool for model comparison. An AUC greater than 0.75 suggests the model may provide useful
predictions. While discrimination is a useful statistic, it is not sufficient to assess model fit and additional measures are needed.

Calibration

Calibration is a more nuanced evaluation of model fit and is crucial for assessing clinical utility of a model. A calibration plot is a visual comparison of predicted versus observed probabilities grouped by percentile of predictions. Models with similar AUC values can have different calibration statistics. Poor calibration on a new dataset can result from unique disease presentation in distinct populations or from overfitting. An overfit model produces a pattern in which risk is overestimated for high-risk cases and underestimated for low ones.

Calibration-in-the-large is a measurement of calibration for the dataset overall and describes the average predicted risk with the observed outcome. This corresponds to the intercept on the calibration plot, which has an ideal value of 0. A negative intercept indicates overestimation of risk overall, while a positive intercept indicates underestimation.

An additional measure of calibration is the slope of the plot. The perfectly calibrated model has a slope of 1, and a model will have perfect calibration on the dataset on which it was developed. Analysis of calibration slope is particularly important when the model is applied to a new dataset. A slope of less than 1 indicates model overfitting (low-risk case predictions are too low and high-risk
predictions are too high). The chance of overfitting can be reduced by employing regularization during model development, as discussed above.

**Internal validation**

To internally validate the model, optimism corrected discrimination and calibration statistics were calculated using the validate function from R-rms, which iteratively applied bootstrap resampling validation of the model for 200 repetitions. In this method, the model is fit to each bootstrap sample and then to the original sample. The optimism is calculated as the difference between model performance on the resample and model performance on the original sample. Optimism corrected statistics were calculated for AUC and slope and intercept of the calibration plot.

**External validation**

The model was developed on data from West Africa and externally validated on a temporally and geographically distinct dataset collected in the DRC. For model validation, we applied the model to the DRC data which comprised 74 complete cases with no missing values. A calibration plot was generated for which we evaluated slope and intercept. An additional measure of goodness-of-fit for logistic regression, the Hosmer-Lemeshow test, was applied to assess if the observed and predicted probabilities match. Similar to generating
a calibration plot, the HL test requires grouping cases by the predicted event probability. The formula for the HL test is:

\[ H = \sum_{j=1}^{10} \frac{(O_j - E_j)^2}{E_j(1 - E_j/n_j)} \sim \chi^2_8 \]

The numerator corresponds to squared difference between observed \( O_j \) and expected \( E_j \) cases for each decile \( j \). The formula produces a chi-square test statistic and a p-value to assess model fit. Additionally, an ROC plot and the associated AUC were calculated to evaluate performance of the model on the DRC data.

**Chapter 2.8 Model Update**

To further improve model performance, we followed the model recalibration with a previously outlined extension protocol.\(^4^9\) We sought to add an additional biomarker as a potentially strong predictor that was available in the external validation data but not in the development data. We focused on commonly used biochemistry laboratory values recorded within 48 hours of admission and selected covariates that were found to be significantly correlated with adverse outcome by Spearman's rank correlation coefficient. We then recalibrated the model and added an additional predictor simultaneously by fitting a new model with the linear predictors of the original model and the additional biomarkers. We did not impute missing data at this step; updated models were fit
only on complete cases. Models were evaluated by comparing their AUCs and 95% confidence intervals.
CHAPTER 3

RESULTS

Chapter 3.1 Summary of West Africa Data

Baseline characteristics

The West Africa dataset included 579 Ebola-positive patients less than 18 years of age with an overall CFR of 40%. Age was among the strongest predictors of mortality with each five-year decrease in age associated with an increase in the odds of death by more than half (Table 3.1; Figure 3.1).

Table 3.1 Demographic and clinical characteristics of patients in the West Africa derivation cohort.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Survived, n=345</th>
<th>Died, n=234</th>
<th>OR (95% CI)(^a)</th>
<th>p-value(^b)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Demographics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (years), median (IQR)</td>
<td>11 (7, 14)</td>
<td>6 (3, 13)</td>
<td>0.55 (0.46 – 0.65)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Male sex, n (%)</td>
<td>159 (46)</td>
<td>112 (48)</td>
<td>1.07 (0.77 – 1.5)</td>
<td>0.67</td>
</tr>
<tr>
<td><strong>Symptoms(^c)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asthenia</td>
<td>297 (88)</td>
<td>171 (74)</td>
<td>0.39 (0.25 – 0.6)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Headache</td>
<td>203 (60)</td>
<td>93 (42)</td>
<td>0.48 (0.34 – 0.68)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Abdominal pain</td>
<td>165 (50)</td>
<td>78 (36)</td>
<td>0.57 (0.4 – 0.81)</td>
<td>0.002</td>
</tr>
<tr>
<td>Bleeding</td>
<td>43 (14)</td>
<td>45 (22)</td>
<td>1.68 (1.06 – 2.67)</td>
<td>0.027</td>
</tr>
<tr>
<td>Joint pain</td>
<td>113 (38)</td>
<td>48 (29)</td>
<td>0.68 (0.45 – 1.01)</td>
<td>0.060</td>
</tr>
<tr>
<td>Bone or muscle pain</td>
<td>121 (37)</td>
<td>61 (29)</td>
<td>0.71 (0.48 – 1.02)</td>
<td>0.067</td>
</tr>
<tr>
<td>Respiratory distress</td>
<td>26 (7.9)</td>
<td>27 (13)</td>
<td>1.69 (0.95 – 2.99)</td>
<td>0.071</td>
</tr>
</tbody>
</table>
A Ct value below 21 was associated with higher mortality at all ages. Variables that were associated with significantly increased odds of survival included asthenia/weakness, headache, and abdominal pain (p < 0.02). In contrast, the presence of bleeding within the first 48 hours of admission increased the odds of death by almost 70% (p < 0.03). While the geographical distribution of cases within West Africa did not reveal a trend in CFR by location (Figure 3.1), there was a modest inverse correlation (r = -0.51) between CFR and number of cases at each ETU, suggesting that patients at treatment centers that had larger numbers of cases may have had less lethal outcomes for reasons that have not been determined. The map in Figure 3.1 shows the geographical distribution of children with EVD included in triage data from the Ebola Data...
Platform, collected during the West African EVD outbreak from 2014-2016. Bubble size corresponds to the number of cases reported, and color corresponds to observed case fatality rate.

**Figure 3.1. Map of ETU locations.**
**Missing data analysis**

In the West Africa data, 10.7% of values were missing from 16 out of 18 predictor variables. Patient age and sex were recorded for all patients. The variable with the highest missingness was CT value, with 47% of values not reported. (Table 3.2).

**Table 3.2 Summary of missing values in West Africa derivation dataset.**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Total missing</th>
<th>% Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ct value</td>
<td>274</td>
<td>47.32</td>
</tr>
<tr>
<td>Conjunctivitis</td>
<td>180</td>
<td>31.09</td>
</tr>
<tr>
<td>Joint pain</td>
<td>115</td>
<td>19.86</td>
</tr>
<tr>
<td>Anorexia</td>
<td>103</td>
<td>17.79</td>
</tr>
<tr>
<td>Nausea</td>
<td>99</td>
<td>17.10</td>
</tr>
<tr>
<td>Any bleeding</td>
<td>66</td>
<td>11.40</td>
</tr>
<tr>
<td>Dysphagia</td>
<td>39</td>
<td>6.74</td>
</tr>
<tr>
<td>Hiccups</td>
<td>39</td>
<td>6.74</td>
</tr>
<tr>
<td>Bone or muscle pain</td>
<td>37</td>
<td>6.39</td>
</tr>
<tr>
<td>Diarrhea</td>
<td>35</td>
<td>6.04</td>
</tr>
<tr>
<td>Respiratory distress</td>
<td>35</td>
<td>6.04</td>
</tr>
<tr>
<td>Abdominal pain</td>
<td>33</td>
<td>5.70</td>
</tr>
<tr>
<td>Headache</td>
<td>24</td>
<td>4.15</td>
</tr>
<tr>
<td>Vomit</td>
<td>18</td>
<td>3.12</td>
</tr>
<tr>
<td>Fever</td>
<td>10</td>
<td>1.73</td>
</tr>
<tr>
<td>Asthenia weakness</td>
<td>9</td>
<td>1.55</td>
</tr>
<tr>
<td>Patient age</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Patient sex</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Multiple imputation results

Multiple imputation was used to address missing data and the protocol was applied to 16 variables with missing values. Figure 3.2 presents a visual depiction of imputed CT values, grouped by clinical presentation of bleeding as a symptom. Consistent with clinical expectations, patients who experienced bleeding have lower real and imputed CT values, indicating a higher viral load and more advanced disease.

Figure 3.2. Distribution of imputed CT values for patients who presented with or without bleeding upon admission.
Chapter 3.2 Summary of DRC Data

Baseline characteristics

The DRC dataset (N = 72) used for model validation has a 48.6% mortality rate. Table 3.3 depicts the baseline characteristics of the dataset.

Table 3.3 Demographic and clinical characteristics of patients in DRC validation cohort.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Survived (n = 37)</th>
<th>Died (n = 35)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Demographics</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male sex, n (%)</td>
<td>20 (54.1)</td>
<td>22 (62.9)</td>
</tr>
<tr>
<td>Age (years), median</td>
<td>7.4</td>
<td>7</td>
</tr>
<tr>
<td><strong>Symptoms</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bleeding n, (%)</td>
<td>1 (2.7)</td>
<td>16 (45.7)</td>
</tr>
<tr>
<td>Diarrhea n, (%)</td>
<td>11 (29.7)</td>
<td>17 (48.5)</td>
</tr>
<tr>
<td>Respiratory distress n, (%)</td>
<td>1 (2.7)</td>
<td>10 (28.6)</td>
</tr>
<tr>
<td>Swallowing problems n, (%)</td>
<td>5 (13.5)</td>
<td>6 (17.1)</td>
</tr>
<tr>
<td>Anorexia n, (%)</td>
<td>21 (56.8)</td>
<td>29 (82.8)</td>
</tr>
<tr>
<td>Bone muscle pain n, (%)</td>
<td>10 (27.0)</td>
<td>14 (40.0)</td>
</tr>
<tr>
<td>Conjunctivitis n, (%)</td>
<td>13 (35.1)</td>
<td>23 (65.7)</td>
</tr>
<tr>
<td>Nausea n, (%)</td>
<td>13 (35.1)</td>
<td>23 (65.7)</td>
</tr>
<tr>
<td>Joint pain n, (%)</td>
<td>9 (24.3)</td>
<td>14 (40.0)</td>
</tr>
<tr>
<td>Abdominal pain n, (%)</td>
<td>14 (37.8)</td>
<td>19 (54.3)</td>
</tr>
<tr>
<td>Headache n, (%)</td>
<td>21 (56.8)</td>
<td>21 (60.0)</td>
</tr>
</tbody>
</table>
Hiccup n, (%) 0 (0) 1 (2.9)
CT, median 24.7 18.2

**Experimental Treatment**

Remdesivir n, (%) 2 (5.4) 3 (8.6)
Regeneron n, (%) 16 (43.2) 8 (22.9)
Mab 114 n, (%) 18 (48.6) 19 (54.3)
ZMAPP n, (%) 0 (0) 1 (2.9)
Missing n, (%) 1 (2.7) 4 (11.4)

A comparison of the derivation and validation cohorts is given in Table 3.4. The validation cohort had a higher mortality rate (46.2% CFR, median CT = 19.3) than the derivation cohort (29.7%, median CT = 25.1), and a generally more severe clinical presentation of disease. Of the four binary symptoms bleeding, diarrhea, respiratory distress, and dysphagia, all were observed more frequently in the validation data. The discrepancy could be partially attributed to the younger median patient age for the validation data (median = 5) than the development data (median = 10), due to more severe disease in children. Additionally, the two datasets are temporally and geographically distinct, and may be associated with different strains of Ebola virus. Thus, the DRC data represents a distinct outbreak and is suitable for external validation of the model.
Table 3.4 Comparison of baseline characteristics in West Africa derivation and DRC validation cohorts

<table>
<thead>
<tr>
<th></th>
<th>Derivation Cohort</th>
<th>Validation Cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case-fatality rate (n, %)</td>
<td>234 (40.4)</td>
<td>35 (46.2)</td>
</tr>
<tr>
<td><strong>Continuous predictors (median, IQR)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>10 (5-14)</td>
<td>5 (1.5-14)</td>
</tr>
<tr>
<td>Ct value</td>
<td>25.1 (20.9-29.5)</td>
<td>19.3 (17.6-26.1)</td>
</tr>
<tr>
<td>**Binary symptoms (n, %)**a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bleeding</td>
<td>88 (15.1)</td>
<td>17 (22.9)</td>
</tr>
<tr>
<td>Diarrhea</td>
<td>57 (9.8)</td>
<td>40 (54.1)</td>
</tr>
<tr>
<td>Respiratory distress</td>
<td>9.7 (1.7)</td>
<td>16 (21.6)</td>
</tr>
<tr>
<td>Dysphagia</td>
<td>19 (3.3)</td>
<td>16 (21.6)</td>
</tr>
</tbody>
</table>

*Covariates presented are those included in the EPiC model. Abbreviations: IQR: interquartile range; Ct: cycle threshold

Chapter 3.3 Model selection and performance

*Variable selection*

The results of the variable selection protocol described in the methods section are given in Table 3.5. Based on the 100% cutoff threshold, four variables were selected for inclusion in the models: bleeding, diarrhea, respiratory distress, and dysphagia (swallowing problems).
Table 3.5 Results of variable selection protocol for categorical variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Percent of models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Any bleeding</td>
<td>1</td>
</tr>
<tr>
<td>Diarrhea</td>
<td>1</td>
</tr>
<tr>
<td>Respiratory distress</td>
<td>1</td>
</tr>
<tr>
<td>Dysphagia</td>
<td>1</td>
</tr>
<tr>
<td>Anorexia</td>
<td>0.97</td>
</tr>
<tr>
<td>Bone or muscle pain</td>
<td>0.64</td>
</tr>
<tr>
<td>Nausea</td>
<td>0.25</td>
</tr>
<tr>
<td>Vomit</td>
<td>0.11</td>
</tr>
<tr>
<td>Conjunctivitis</td>
<td>0.1</td>
</tr>
<tr>
<td>Joint pain</td>
<td>0.07</td>
</tr>
<tr>
<td>Hiccups</td>
<td>0.05</td>
</tr>
<tr>
<td>Fever</td>
<td>0.03</td>
</tr>
<tr>
<td>Patient sex</td>
<td>0.01</td>
</tr>
<tr>
<td>Abdominal pain</td>
<td>0</td>
</tr>
<tr>
<td>Asthenia/weakness</td>
<td>0</td>
</tr>
<tr>
<td>Headache</td>
<td>0</td>
</tr>
</tbody>
</table>

Model derivation

The initial model included two continuous predictors (age and CT value) and four binary covariates (bleeding, diarrhea, respiratory distress, dysphagia). Internal validation using bootstrap indicated strong performance with an optimism-corrected AUC of 0.77 (95% CI: 0.74-0.81).

External validation

The AUC on the external validation DRC data was 0.76 (95% CI: 0.64-0.88) (Figure 3.3). In the discrimination plot, the receiver operating characteristic (ROC) curve is plotted (central black line) together with the 95% confidence
interval band (blue shaded area). In the calibration plot, the dots represent the mean estimate of the observed probability for each 10% bin of predicted probability (with probability being risk of death), the vertical lines passing through each dot are the corresponding confidence intervals for the observed probability, the dashed line is the best linear fit passing through the mean values, and the red line is the LOESS curve fitting all the individual observed/predicted pairs in the data.

**Figure 3.3 Performance characteristics of the prediction model.** Discrimination (A) and calibration (B) plots of the model are shown for the Democratic Republic of the Congo validation dataset.

To quantitatively assess the predictive value of the EPiC model, we considered the slope and intercept of the linear fit of the calibration data and compared it against the ideal: a slope of 1 and an intercept of 0 (Figure 3.3b).
The slope was 0.89 and the intercept -0.09, indicating that the EPiC model provides a good risk estimation overall, with only a small bias towards overestimating risk of death for all patients (except for one outlier point corresponding to a single high-risk patient) by approximately 0.1 on average. The confusion matrix and additional performance measures calculated at the optimal prediction cutoff for accuracy ($p_{\text{cutoff}} = 0.63$) are presented in Tables 3.6a & 3.6b. The $p_{\text{cutoff}}$ value of 0.63 is consistent with the bias of around 0.1 towards risk overestimation observed in the calibration plot.

**Tables 3.6a & 3.6b Confusion matrix and detailed performance measures of the EPiC model.** Statistics reflect model as applied on the validation DRC dataset at the optimal prediction cutoff for accuracy (0.63). Both tables were generated with the confusionMatrix() function in the R package Caret.50

**Table 3.6a**

<table>
<thead>
<tr>
<th>Predicted</th>
<th>Observed</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Survived</td>
<td>Died</td>
</tr>
<tr>
<td>Survived</td>
<td>25</td>
<td>5</td>
</tr>
<tr>
<td>Died</td>
<td>11</td>
<td>26</td>
</tr>
</tbody>
</table>

**Table 3.6b**

- **Accuracy**: 0.76
- **Accuracy 95% CI**: (0.64, 0.87)
- **Sensitivity**: 0.84
- **Specificity**: 0.69
- **Pos Pred Value**: 0.70
Neg Pred Value 0.83
Prevalence 0.46
Detection Rate 0.39
Detection Prevalence 0.55
Balanced Accuracy 0.77
False Alarm Rate 0.3
Miss Rate 0.16

Comparison with previously published models

Prior prognostic models\textsuperscript{51-54} are not specific to pediatric patients and may use different features, which makes comparisons difficult. To compare performance of our model to a previously published model, we applied the minimal (age+CT) model from,\textsuperscript{51} which was trained on all patients (pediatric and adult) from the IMC ETUs in the West African EVD outbreak (a subset of the EDP dataset), on the DRC dataset. The performance is shown in Figure 3.4, which shows a similar AUC of 0.77 (95% CI: 0.65-0.88), but a poorly calibrated model, with a slope of 1.22 and an intercept of -0.27 in the linear fit to the calibration data.
Figure 3.4 Receiver operating characteristic (ROC) (A) and calibration (B) plots for the minimal model (age+CT) described previously. Model was trained on all patients (not pediatric-specific) in the EVD West African dataset from IMC. The intercept and slope of the linear fit to the predicted probabilities are -0.27 and 1.22, respectively.

This is consistent with previous observations that clinical features make a small contribution to the prediction relative to age and viral load, as it can be seen for our model in the ANOVA and odd ratios (OR) charts in Figure 3.5. In the figure, the analysis of variance chart was generated with the anova function in the rms package, showing a ranking of the features according to their predictive contribution to the model, as measured by the Wald χ2-d.f. (degrees of freedom) statistic. Although the effect is small, inclusion of selected clinical features
consistently results in better calibrated models, for both our own and previously published models.

**Figure 3.5 Plots showing the importance of the features in the EPiC model.**

![Figure 3.5 Plots showing the importance of the features in the EPiC model.](image)

**Chapter 3.4 Model Update**

We sought to improve model performance by recalibrating the intercept and slope of the calibration plot and adding a biomarker to the model that was only available in the DRC data. An analysis of peak laboratory test results measured within the first 48 hours after admission identified three variables each significantly (p <0.01) correlated with mortality: ALT (r=0.57), AST (r=0.56), and CK (r=0.51). We omitted ALT because it is highly colinear with AST (Pearson
correlation = 0.83). Despite limited availability of test results in the validation data (AST: n=29; CK: n=33), we used these new variables to build additional models.

Models that incorporated an additional predictor outperformed the original EPiC model on the validation data, in which adding CK as a predictor produced an AUC of 0.87 (95% CI: 0.74-1) while adding AST gave an AUC of 0.90 (95% CI: 0.77-1). We also considered a third model with both AST and CK added as predictors, since the association between these two biomarkers was moderate (Pearson correlation = 0.52), suggesting that they contain some amount of mutually independent information that could be combined to improve the predictions. Indeed, the model with AST and CK yields a higher AUC of 0.95 (95% CI: 0.86-1). The confusion matrix for this model exhibits an almost perfect discriminative capability with only 1 misclassification in each outcome category (Tables 3.7a & 3.7b). However, the sample size for this model was reduced further to n=23, since it requires patients to have data for both biomarkers. The ROCs and calibration plots for these three models are shown in Figure 3.6.
Tables 3.7a & 3.7b Confusion matrix and detailed performance measures of the EPiC model augmented with AST and CK. Model was applied on the validation DRC dataset at the optimal prediction cutoff for accuracy (0.5). Interpretation of these tables is the same as in 3.6a & 3.6b.

Table 3.7a

<table>
<thead>
<tr>
<th></th>
<th>Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Survived</td>
</tr>
<tr>
<td>Predicted</td>
<td></td>
</tr>
<tr>
<td>Survived</td>
<td>16</td>
</tr>
<tr>
<td>Died</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 3.7b

<table>
<thead>
<tr>
<th>Measure</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.91</td>
</tr>
<tr>
<td>Accuracy 95% CI</td>
<td>(0.72, 0.99)</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.83</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.94</td>
</tr>
<tr>
<td>Pos Pred Value</td>
<td>0.83</td>
</tr>
<tr>
<td>Neg Pred Value</td>
<td>0.94</td>
</tr>
<tr>
<td>Prevalence</td>
<td>0.26</td>
</tr>
<tr>
<td>Detection Rate</td>
<td>0.22</td>
</tr>
<tr>
<td>Detection Prevalence</td>
<td>0.26</td>
</tr>
<tr>
<td>Balanced Accuracy</td>
<td>0.89</td>
</tr>
<tr>
<td>False Alarm Rate</td>
<td>0.06</td>
</tr>
<tr>
<td>Miss Rate</td>
<td>0.17</td>
</tr>
</tbody>
</table>
Figure 3.6 Discrimination and calibration curves for model with additional biomarker. Area under the receiver operating characteristic curves (AUC) (A, C, E) and calibration curves (B, D, F) of the Ebola Virus Disease Prognosis in Children (EPiC) model are shown with aspartate aminotransferase (AST) (A, B), creatine kinase (CK) (C, D), or both (E, F) as additional predictors for the Democratic Republic of the Congo validation dataset. The interpretation of the plots is the same as in Figure 3.4.
CHAPTER 4: DISCUSSION

Chapter 4.1 Overview of contribution to the field

This study presents a newly derived and externally validated prognostic model for pediatric EVD. The model shows that younger age, lower Ct values, and bleeding are associated with a poor prognosis, while asthenia, headache and abdominal pain predict better outcomes. Predicting survival for children with EVD is uniquely challenging because the epidemiology and complications of EVD vary between outbreaks due to differing health seeking behaviors, viral dynamics, medical interventions, and socioeconomic, cultural, and political contexts.\textsuperscript{33,55}

Updating the EPiC model with additional biomarkers (AST, or CK) substantially improved model performance. Elevated levels of AST have been observed in patients with EVD and are associated with more severe, potentially fatal disease.\textsuperscript{52,56,57} Elevated AST likely reflects not only viral-induced hepatitis but also damage to other cells and end-organs such as red blood cells, pancreas, muscle, or kidneys. To our knowledge, CK has not been previously described as a predictive biomarker for EVD outcomes. Such biomarker data can be useful in helping to predict mortality for pediatric patients with EVD. For instance, shock may lead to an increase in AST/ALT and CK. However, the results must be interpreted with caution due to the small sample size.

The model building approach was based on Elastic Net, a form of regularized regression that has performed favorably against stepwise
Regression.\textsuperscript{58,59} Regularized regression is particularly good at retaining explanatory variables while reducing model complexity by removing nuisance variables. The final EPiC model that emerged from the Elastic Net-based variable selection protocol is parsimonious in its complexity and the included predictors of EVD severity match clinical intuition. Furthermore, this protocol was easily extended to update the model with additional biochemical predictors available in the DRC data. These compelling results suggest that our variable selection and model update protocol could be applied to other similar datasets.

Chapter 4.2 Limitations of research

A limitation of this study is the moderate amount of missing data (approximately 10\%) for some variables, which highlights the difficulty of collecting data during a humanitarian emergency. Also, some patients may have been given experimental treatment under compassionate use, but such detailed information is not available in the West Africa derivation dataset. Furthermore, only aggregated data by day was available. As such, it was not possible to determine whether a patient died immediately upon arrival or later that day, requiring us to exclude all patients who died with one day of admission from our prediction model. The good overall calibration of our model suggests that such exclusion did not significantly affect the predictions.

The derivation dataset was collected from several different humanitarian agencies with differing data collection and laboratory procedures. Therefore, the
scale of Ct values may vary between various laboratories. All Ct values presented in this manuscript were used to derive, validate, and update the models without any sort of normalization to account for the potential differences across the laboratories in the EDP dataset. Rerunning the calculations with normalized Ct values (obtained by subtracting the mean and dividing by the standard deviation at each site) revealed that all AUC values remained the same except for the AUC value on the validation dataset which was slightly lowered from 0.76 (CI 0.64-0.88) to 0.71 (CI 0.59-0.84). This indicates that the effect of Ct differences across sites is not large, but also that models could be improved if raw Ct data were more consistent, or a more rigorous inter-site normalization protocol could be defined.

The validation dataset is small (74 cases total) due to inclusion of only those cases with complete data, so study results have to be interpreted with caution, particularly those from model updating, which further reduced the sample size. However, these favorable preliminary results provide compelling justification for future prospective studies to investigate the prognostic utility of certain biomarkers for children as well as adults. These biomarkers, which are often part of a standard blood chemistry panel, are more accessible in low resource settings than more expensive testing such as proinflammatory cytokines. Furthermore, collecting symptom information from children is difficult, especially from those who have not developed verbal skills. In fact, upon further testing, we found bone and muscle pain, asthenia, headache, and abdominal
pain to be correlated with age, illustrating that children in the pre-verbal age group (defined as <2 years of age) cannot reliably report these symptoms. Lastly, both settings adhered to WHO treatment guidelines and each country’s respective national guidelines. As such, there may have been slight differences in the treatment protocol between the West Africa derivation cohort and the DRC validation cohort.

Chapter 4.3 Conclusion

The EPiC model is the first externally validated model for the prognosis of pediatric EVD. Pediatric patients with asthenia/weakness, headache, and abdominal pain were more likely to survive, while younger children or those with lower Ct values, bleeding, diarrhea, respiratory distress, and/or dysphagia were more likely to die from EVD. As Ct value is a strong clinical predictor, rapid molecular tests should be widely available. The addition of routine blood test biochemical markers, such as AST and CK, strengthened the model and are usually available. This model can be easily applied by bedside clinicians to assess pediatric patients at risk for death and help to allocate resources accordingly. An online calculator has been developed so that clinicians can conveniently use the EPiC model to calculate risk scores, available at: https://kelseymbutler.shinyapps.io/epic-calculator/. Future improvements of this model would result from larger sample sizes with more consistent variable definitions and protocols across sites.
REFERENCES


31 Smit MA, Michelow IC, Glavis-Bloom J, Wolfman V, Levine AC. Characteristics and Outcomes of Pediatric Patients With Ebola Virus Disease Admitted to Treatment Units in Liberia and Sierra Leone: A Retrospective Cohort Study. *Clinical Infectious Diseases* 2017; 64: 243–9.


40 Harrell F, Dupont C. Hmisc: Harrell Miscellaneous.


42 Friedman J, Hastie T, Tibshirani R. glmnet: Lasso and elastic-net regularized generalized linear models.


45 Harrell F. rms: Regression Modeling Strategies.


Kuhn M, Wing J, Weston S. caret: Classification and Regression Training. 


APPENDIX A

DATA COLLECTION METHODOLOGY

Study Design and Setting

This study used retrospective data from children presenting to Ebola treatment units (ETUs) in West Africa and the DRC. The West Africa derivation dataset was built from the Infectious Diseases Data Observatory’s (IDDO) Ebola Data Platform (EDP). IDDO’s EDP is the first global data repository for clinical, epidemiological, and laboratory data from patients with EVD during the 2014 - 2015 West Africa outbreak (specifically Liberia, Guinea and Sierra Leone) provided by the following organizations: Alliance for International Medical Action (ALIMA), International Medical Corps (IMC), Institute of Tropical Medicine Antwerp (ITM), Médecins Sans Frontières (MSF), University of Oxford, Save the Children International (SCI), who had no role in the conduction of this study.

The validation DRC dataset was derived from patients who presented at IMC’s Mangina ETU during the 2018 - 2020 EVD outbreak in the DRC. The DRC’s eastern provinces of North Kivu and Ituri served as the main catchment area for the Mangina ETU, located in North Kivu.

Participant Selection

All patients less than 18 years of age who presented to West African ETUs from June 2014 to October 2015 and to the Mangina ETU from December 2018 to January 2020 with laboratory confirmed EVD were eligible for inclusion in
the derivation and validation datasets, respectively. Patients were excluded if they had missing outcome data or if they died on the day of admission to the ETU.

_EVD Triage and Diagnosis_

**West Africa**

Since the data from Liberia, Guinea, and Sierra Leone were provided by several humanitarian aid organizations, triage procedures varied slightly from site to site. All organizations adhered to World Health Organization (WHO) diagnostic criteria and relevant national guidelines.

**DRC**

All patients presenting at the IMC’s Mangina ETU were screened by trained clinical staff to ensure they met the clinical case definition for suspected EVD based on WHO and MSF guidelines and in consultation with local health authorities. If patients presented with a documented diagnosis of EVD, they were directly admitted to the ward for patients with confirmed disease. Otherwise, patients who met the case definition but had no prior testing were admitted to the ETU’s ward for suspected cases, where they underwent EVD testing. If the patient’s initial test was negative, they remained in the ETU until 72 hours had passed, and a second EVD test was negative, in which case they were discharged. Patients with a positive test result were moved to the “confirmed” ward for further management.
**Laboratory Methods**

All PCR cycle threshold (Ct) values presented in this study are based on RT-PCR of the same Zaire *ebolavirus* nucleoprotein locus using standardized RNA extraction procedures. A Ct greater than 40 was considered negative in all cases.

**West Africa**

Data were provided by several humanitarian aid organizations and consequently laboratory methods differed slightly among treatment sites.

**DRC**

DRC’s ETUs received all patients from the surrounding catchment areas some of whom may or may not have had laboratory confirmed EVD in the community or other test facility prior to arrival. Patients were diagnosed with EVD with a RT-PCR (GeneXpert) blood assay using plasma. Blood chemistry tests were completed at point of care using Piccolo Amlyte 13, which determined levels of glucose, creatinine (CRE), albumin (ALB), aspartate aminotransferase (AST), alanine aminotransferase (ALT), amylase (AMY), potassium, C-reactive protein (CRP), total urea nitrogen (BUN), total bilirubin (TBL), creatine kinase (CK), sodium, and calcium.